

The genetic factors of bilaterian evolution.

Peter Heger^{1*}, Wen Zheng^{1‡}, Anna Rottmann¹, Kristen A. Panfilio^{3,4}, Thomas Wiehe¹

*For correspondence:

peter.heger@uni-koeln.de (FMS)

Present address: [‡]West

China-Washington Mitochondria and Metabolism Research Center, West China Hospital, Sichuan University, Chengdu, China

¹ Institute for Genetics, Cologne Biocenter, University of Cologne, Zùlpicher Straße 47a, 50674 Köln, Germany; ³ Institute for Zoology: Developmental Biology, Cologne Biocenter, Zùlpicher Straße 47b, 50674 Köln, Germany; ⁴ School of Life Sciences, University of Warwick, Gibbet Hill Campus, Coventry CV4 7AL, UK

Abstract The Cambrian explosion was a unique animal radiation ~540 million years ago that produced the full range of body plans across bilaterians. The genetic mechanisms underlying these events are unknown, leaving a fundamental question in evolutionary biology unanswered. Using large-scale comparative genomics and advanced orthology evaluation techniques, we identified 157 bilaterian-specific genes. They include the entire Nodal pathway, a key regulator of mesoderm development and left-right axis specification; components for nervous system development, including a suite of G protein-coupled receptors that control physiology and behaviour, the Robo-Slit midline repulsion system, and the neurotrophin signalling system; a high number of zinc finger transcription factors; and novel factors that previously escaped attention. Contradicting the current view, our study reveals that genes with bilaterian origin are robustly associated with key features in extant bilaterians, suggesting a causal relationship.

Introduction

The taxon Bilateria consists of multicellular animals with bilateral body symmetry and constitutes a major and ancient radiation of animals. There is compelling morphological and molecular evidence for the monophyly of bilaterians (Hejnol *et al.*, 2009; Dunn *et al.*, 2014; Cannon *et al.*, 2016), for their subdivision into protostomes and deuterostomes (Aguinaldo *et al.*, 1997; Philippe *et al.*, 2005; Dunn *et al.*, 2008; Simakov *et al.*, 2013; Cannon *et al.*, 2016), and for the overall relationships of ~25 phyla that make up this group (Dunn *et al.*, 2008; Hejnol *et al.*, 2009; Dunn *et al.*, 2014). In contrast, the evolutionary relationships of non-bilaterian metazoans are still a matter of debate, in particular the relative positions of placozoans, ctenophores, and sponges (Brooke and Holland, 2003; Ryan *et al.*, 2013; Pisani *et al.*, 2015; Feuda *et al.*, 2017; Simion *et al.*, 2017; Whelan *et al.*, 2017).

The first unambiguously bilaterian fossils appear in Cambrian sediments with an age of ~540 million years (Marshall, 2006; Erwin and Valentine, 2013). By the end of Cambrian stage 3 (499 Mya), stem groups of all major bilaterian phyla inhabited Earth. This abrupt appearance of most bilaterian body plans, the sets of morphological features common to a phylum, already puzzled Darwin (Darwin, 2009). It is considered one of the most important evolutionary events after the origin of life (Conway Morris, 2006; Budd, 2008) and still awaits an explanation today. Importantly, no new body plans evolved in the 500 My since the initial radiation.

Abiotic, ecological, and genetic factors have been proposed to explain the Cambrian radiation. While deep-ocean oxygenation (Canfield *et al.*, 2007), the availability of calcium (Jackson *et al.*, 2010), or ecological interactions (Budd and Jensen, 2017) likely played a role, genetic changes in

the bilaterian ancestor must ultimately have constituted its molecular basis. However, evidence for such genetic changes is scarce. Genomic sequencing of non-bilaterian animals revealed that the major signalling pathways and many developmentally important genes of bilaterians are also present in non-bilaterians, indicating that these genes evolved before the advent of bilaterians (Technau et al., 2005; Putnam et al., 2007; Srivastava et al., 2008, 2010; Ryan et al., 2013; Babonis and Martindale, 2017). Similarly, epigenetic mechanisms to regulate gene expression, such as DNA methylation and histone modifications, seem to be conserved between bilaterians and non-bilaterian metazoans (Zemach et al., 2010; Schwaiger et al., 2014). Therefore the common view is that modification of existing gene regulatory networks rather than the invention of new genes determined the evolution of complex body plans (Davidson and Erwin, 2006; Su and Yu, 2017).

Nevertheless, a number of studies identified genes that emerged in the ancestor of bilaterians. One example is a major expansion of miRNA families that likely triggered an increase in miRNA-mediated gene regulation (Prochnik et al., 2007; Wheeler et al., 2009). However, the significance of this event at the base of the Bilateria is unclear because frequent miRNA expansions are seen in various lineages over time (Peterson et al., 2009). Similarly, a link between the genome organizer CTCF and Hox genes presumably emerged in the bilaterian ancestor and might have contributed to the organization of bilaterian body plans (Heger et al., 2012). The importance of CTCF for Hox gene expression has been shown repeatedly (Mohan et al., 2007; Kim et al., 2011; Rousseau et al., 2014; Narendra et al., 2015), yet direct evidence for the involvement of a Hox-CTCF link in body patterning is lacking. Another study implicated the TATA-box-binding protein-related factor 2 (TRF2) in the evolution of bilaterians. This factor may have founded new, TATA box-independent transcriptional programs involved in body plan development (Duttke et al., 2014), but the consequences of this hypothesis have not been tested.

Therefore a comprehensive screen for bilaterian-specific genes and an assessment of their evolutionary impact is missing. A major obstacle for such a screen is the uneven coverage of the animal tree with sequence data. While some lineages, particularly those including model organisms (e.g. nematodes, flies, or mammals), are well represented, other areas of the metazoan tree are remarkably under-represented, e.g. lophotrochozoans and non-bilaterian metazoans. For example, the leading orthology databases OrthoDB (Kriventseva et al., 2015), eggNOG (Huerta-Cepas et al., 2016), and OrthoMCL (Li et al., 2003) contain only two to four non-bilaterian species, and two of these databases do not contain lophotrochozoans at all (Figure 1, Table 1). It is therefore difficult to deduce from such databases the genes that are widespread in bilaterians and absent in non-bilaterians. In addition to the bias in coverage, sequence databases suffer from annotation errors, which particularly affect non-model organisms and under-represented parts of the tree, such as non-bilaterian metazoans and lophotrochozoans. Annotation errors, in turn, have been found as the largest single source for errors in orthology benchmark testing and, together with uneven phylogenetic coverage, accounted for up to 40 % of incorrect assignments (Trachana et al., 2011).

To address these biases and to infer bilaterian-specific genes in a reliable and robust way, we (i) assembled a dataset covering the animal tree in the most comprehensive and representative way so far; (ii) particularly strengthened resolution at the base of the Bilateria; (iii) reduced annotation errors by incorporating newly generated ORF (open reading frame) data sets; and (iv) evaluated the composition of the generated orthologous groups in a phylogenetic context. Using this strategy we extracted, from an initial set of 124 million sequences from 273 species, 157 high-confidence bilaterian-specific genes, with many functions connected to key bilaterian features.

Results

Dataset generation and orthogroup evaluation

Non-bilaterian metazoans are severely under-represented in existing sequence collections, but sufficient coverage is critical to illuminate bilaterian evolution. To maximise phylogenetic resolution at the origin of Bilateria, we assembled a new database specifically tailored to this purpose, the

Table 1. Comparison of three major orthology databases with the BigWenDB. The number of species of a given taxon (left column) in four different orthology databases is shown. In contrast to other databases, the BigWenDB has substantially more sequence information from non-bilaterian metazoans and therefore a better resolution at the divergence of bilaterians and non-bilaterians. D = Deuterostomia, E = Ecdysozoa. Note the bias of other databases towards insects and vertebrates.

Taxon	OrthoDB V8	eggNOG V4.5	OrthoMCL V5	BigWenDB
Cellular organisms	3027	2031	150	273
Metazoa	173	88	29	175
Bilateria	169	85	27	142
non-Bilateria	4	3	2	33
Ecdysozoa (E)	97	29	12	54
E w/o insects	17	9	4	29
Lophotrochozoa	5	0	0	18
Deuterostomia (D)	66	55	14	65
D w/o vertebrates	5	4	1	12

BigWenDB (**Figure 1, Figure 1–Figure Supplement 1; Table 1**). This database combines sequence data of 273 species from three sources. The backbone of our analysis is the opisthokont sequence space (primarily fungi, vertebrates, and insects): 204 species, each with >8,000 available sequences at GenBank, totalling 2.7 million sequences (**Table 2**; NCBI GenBank release 203 from August 15, 2014). The second part derives from transcriptome sequences of 64 species from various sources (Supplementary File 1–Supplementary Table 1, Supplementary File 1–Supplementary Table 2, Supplementary File 2). Among others, non-bilaterian metazoans (30 species) and lophotrochozoans (12 species) contribute 11.7 million sequences to this group, complementing their poor GenBank representation (**Figure 1–Figure Supplement 1**). The third and largest sequence set contains ~109 million open reading frames (ORFs) obtained by translating 25 metazoan genomes (Supplementary File 1–Supplementary Table 3). All non-bilaterian and lophotrochozoan whole genome sequences available at the time, as well as genomes from additional phyla, were included to compile a comprehensive and representative dataset (**Figure 1–Figure Supplement 1**). As this strategy involved a large increase in sequence number, we limited the third set to 25 species to maintain technical feasibility. The final dataset combines 124 million sequences from 21 metazoan and three outgroup phyla, including several taxa absent from other databases, e.g. tardigrades, a priapulid, bryozoans, a nemertean, a rotifer, a brachiopod, and choanoflagellates (**Figure 1, Figure 1–Figure Supplement 1**).

To be able to generate clusters of orthologous proteins from this large dataset, we adapted the OrthoMCL pipeline (**Li et al., 2003**) and improved its scalability (see Appendix 1: Orthology pipeline and clustering; Supplementary File 1–Supplementary Table 4). As a large proportion of the resulting 824,605 orthogroups was small and had phylogenetically inconsistent composition (Appendix 1–Figure 1; Supplementary File 1–Supplementary Table 5), we focused our analysis on 75,744 orthogroups (OGs) with at least ten species. They provide a rich repertoire for the identification of lineage-specific protein sets.

Hundreds to thousands of novel translated open reading frames exist in humans and other animals, that are missed by traditional annotation methods (**Ladoukakis et al., 2011; Mackowiak et al., 2015; Raj et al., 2016**). A key aspect of our analysis is therefore the inclusion of genomic ORFs. To estimate their contribution to the clustering process, we examined the composition of all orthogroups. Genomic ORFs constitute a substantial fraction of the majority of orthogroups, comprising >90 % of all sequences in 50 % of orthogroups. This demonstrates that a high percentage of orthogroups is either dependent on or substantially affected by the inclusion of ORFs. Although most ORFs are short (mean length of 60 AA; **Figure 1–Figure Supplement 2, Figure 1–Figure Supplement 3**), nearly 2.3 million ORFs (on average 90,443 per species) are >132 AA, the mean size of

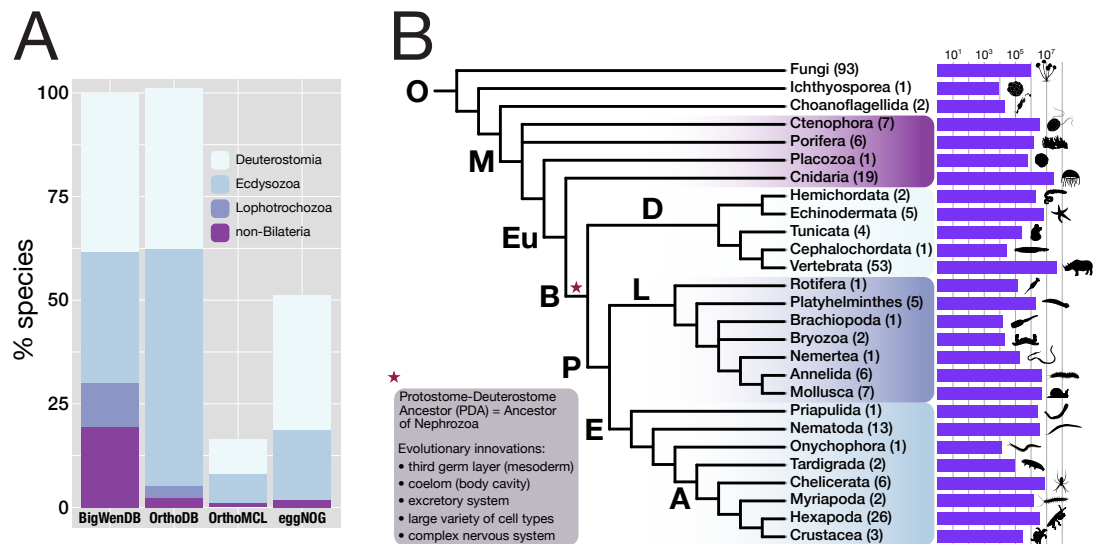


Figure 1. Properties of the BigWenDB data collection. A: Comparison of three major orthology databases with the BigWenDB. The relative contribution of four metazoan clades (Deuterostomia, Ecdysozoa, Lophotrochozoa, and the paraphyletic group «non-Bilateria») is shown as stacked bar graph. The count of metazoans in our database (175 species) is set to 100 %. In comparison to other databases, the BigWenDB has a larger repertoire of critical lophotrochozoans and non-bilaterian Metazoa. **B:** Consensus phylogeny describing the relationships of 21 metazoan phyla covered in our database, after *Laumer et al. (2015)*; *Telford et al. (2015)*; *Torruella et al. (2015)*; *Cannon et al. (2016)*. Bold labels to the left or above branches indicate its ancestor (A: Arthropoda, B: Bilateria, D: Deuterostomia, E: Ecdysozoa, Eu: Eumetazoa, L: Lophotrochozoa, M: Metazoa, O: Opisthokonta, P: Protostomia). Numbers in parentheses (after the phylum name) indicate the number of species present from this phylum. Horizontal bars visualize the number of database sequences that belong to a given phylum (logarithmic scale; transcriptomic, ORF, and NCBI sequences summed up). Species silhouettes were downloaded from www.phylopic.org. Morphological innovations of Bilateria according to *Baguña et al. (2008)* are highlighted in a shaded box.

Figure 1-Figure supplement 1. Phylogenetic distribution of the BigWenDB.

Figure 1-Figure supplement 2. Size distribution of three sequence data types present in the BigWenDB.

Figure 1-Figure supplement 3. ORF size distribution for 25 species with genomic data.

domains in the PFAM database, ensuring the possibility of annotating ORF-dominated orthogroups (Figure 1-Figure Supplement 2).

We next assessed the accuracy and biological validity of our orthogroup dataset via several approaches. First, we compared our clustering results with an external benchmark set of 70 manually curated orthogroups (*Trachana et al. (2011)*; see Appendix 1: Cluster evaluation and quality control; Supplementary File 3). We then specifically examined the clustering results of a highly conserved and difficult to assess class of proteins, the Nkx homeodomain proteins (Supplementary File 1-Supplementary Table 6). Third, we evaluated potential sources of error with respect to the phylogenetic composition of a given orthogroup (see Appendix 1: Identification of bilaterian-specific genes). For this purpose, we developed a new reciprocal HMM-HMM comparison step. It performs sensitive, BLAST-independent searches for orthogroups with similar sequence profiles to validate orthogroup completeness. We demonstrated the value of this step by using two proteins as test cases, the FGF signalling pathway component Sprouty and the insulator protein GAGA factor (see Appendix 1: Identification of bilaterian-specific genes; Supplementary File 1-Supplementary Table 7). After these quality control steps, we finally identified 157 orthogroups as a minimal set of high confidence, bilaterian-specific orthogroups (Supplementary File 4).

The domain repertoire of bilaterian-specific proteins is enriched for DNA-binding

To reveal the putative function of the 157 identified bilaterian-specific genes, we first determined their protein domain repertoire and the gene ontology terms for molecular function associated

Table 2. Composition of the BigWenDB. The number of sequences (overall: 124,031,501) collected from three different sources (NCBI, Transcriptome, ORFs) is indicated for major taxonomic groups of the BigWenDB. «Others» comprises the ichthyosporean *Capsaspora owczaraki* and the choanoflagellates *Monosiga brevicollis* and *Salpingoeca rosetta*.

Group	(Super)Phylum	# Species	NCBI	Transcriptome	ORFs
Bilateria	Deuterostomia	65	895,084	2,292,541	51,922,654
	Ecdysozoa	54	511,663	2,150,424	17,338,026
	Lophotrochozoa	23	170,379	2,618,518	9,805,405
Non-Bilat.	Ctenophora	7	0	1,468,372	2,458,546
	Placozoa	1	11,215	0	590,820
	Porifera	6	8,836	539,299	1,008,535
	Cnidaria	19	36,873	2,361,032	26,443,358
Fungi		93	1,032,299	0	0
others		3	29,292	0	0
total		273	2,695,641	11,768,516	109,567,344

with these domains. We then compared the results to analyses carried out for the vertebrate and arthropod nodes, as these nodes represent major radiations that are well-supported by genome sequence data. The obtained terms indicate that membrane processes, including cell adhesion, G protein-coupled receptor signalling, and Ca^{2+} -binding, as well as protein interactions and metal ion binding, are prominent molecular functions of bilaterian-specific proteins (**Figure 2** left, top and middle row). In contrast, terms derived from the arthropod and vertebrate nodes are markedly different. While the vertebrate repertoire comprises G protein-coupled receptors, cadherins, and extracellular domains required for protein-protein or protein-ligand interactions, arthropod-specific genes are characterised by a broad spectrum of similarly prominent functions, from expected roles in cuticle and chitin biology to a plenitude of conserved domains of unknown function (**Figure 2** middle and right, top and middle row). These results indicate that proteins with distinct functions characterize the evolution of each of the three nodes.

Further, our comparative analysis implied that a large number of transcription factors emerged in the bilaterian ancestor. While 3.58 % of vertebrate-specific orthogroups and 9.30 % of arthropod-specific orthogroups had transcription factor-associated domains such as zinc fingers or homeodomains, the corresponding fraction was 26.06 % in bilaterian-specific orthogroups (**Figure 2** middle row). To substantiate this result, we randomly selected ten times 157 proteins from a curated set of 20,205 human proteins. The average number of transcription factors in these control sets was 12.8 ± 4.44 as opposed to 37 transcription factors in the set of 157 bilaterian-specific genes. This is a highly significant result under a number of assumptions for data distribution (see Methods), lending statistical support to an unexpectedly high number of transcription factors in the bilaterian-specific dataset.

Importantly, many of the transcription factors contained tandem C_2H_2 zinc finger domains and already originated with multiple zinc fingers, as their extant *Drosophila* and human orthologs suggest (Supplementary File 1–Supplementary Table 8). With the addition of at least 13 members, the modest poly-ZF repertoire at the dawn of metazoans thus almost doubled in the bilaterian ancestor (**Figure 2–Figure Supplement 1**), in line with previous evidence that poly-ZF proteins emerged from a small group of eukaryotic zinc finger transcription factors (**Emerson and Thomas, 2009**). Considering that several factors with this domain configuration are involved in regulating chromatin architecture, including CTCF (**Phillips-Cremins et al., 2013**), YY1 (**Weintraub et al., 2017**), Pita (**Kyrchanova et al., 2017**), SuHw (**Van Bortle et al., 2012**), and Casz1 (**Mattar et al., 2018**), these findings open the possibility that multiple poly-ZF factors participated in modifying higher-order chromatin

structure during the emergence of bilaterians, as proposed for CTCF (Heger et al., 2012; Vietri Rudan and Hadjur, 2015; Acemel et al., 2017). With the exception of YY1 (OG_3966: metazoan origin or earlier), all known chromatin architectural proteins emerged in the ancestor of bilaterians or later (Heger et al., 2013; Heger and Wiehe, 2014), suggesting that a more sophisticated regulation of gene expression by influencing chromatin architecture contributed to bilaterian evolution. More generally, we note that poly-ZF proteins often comprise the most abundant transcription factor superfamily in bilaterians, with many lineage-specific expansions even within orders and families (Panfilio et al., 2019). Below, we also comment both on similar patterns in other protein classes and on potential other roles of a bilaterian expansion in poly-ZF proteins.

Bilaterian-specific proteins contain novel protein domains

Using domain scans, we could not identify known protein domains or other functional annotation for five of the 157 bilaterian-specific orthogroups. Nevertheless, the corresponding alignments displayed extended regions of sequence conservation (Figure 2–Figure Supplement 2, Figure 2–Figure Supplement 3, Figure 2–Figure Supplement 4), arguing that these regions may constitute so far undetected protein domains. To explore whether the putative domains are bilaterian novelties, we converted them to hidden Markov models and used these to search our database of 824,605 orthogroup HMMs. In these searches, only one of the five domains showed weak evidence for homology outside the Bilateria, indicating that a protein with a similar domain exists in non-bilaterians. The other four domains were restricted to bilaterians, like the proteins they belong to (Supplementary File 1–Supplementary Table 9), a finding compatible with the *de novo* birth of these five genes. Similarly, sequences without known protein domains were also detectable in arthropod- and vertebrate-specific orthogroups (Figure 2) and, more generally, in approximately 40 % of the 69,114 orthogroups with more than ten species. These findings open the possibility that, across opisthokonts, many lineage-specific genes are uncharacterised and may contain previously undescribed protein domains and novel lineage-specific domains, emphasizing the involvement of gene birth in lineage evolution on a broad scale.

Changes in the transcription factor repertoire and in membrane processes accompany bilaterian evolution

Nuclear factors include key developmental regulators

To reveal the putative function of the identified bilaterian-specific genes, we determined the sub-cellular location of their human orthologs according to the information at www.uniprot.org (Figure 3). Almost two-thirds of the 157 genes belonged to either of two cellular compartments, the nucleus or the plasma membrane. The majority of nuclear proteins (40/57 orthogroups) had transcription factor activity, with various domains for DNA binding (Figure 3B). Although C₂H₂ poly-ZF proteins are particularly enriched (Figure 2–Figure Supplement 1, Supplementary File 1–Supplementary Table 8), we also found several transcription factors with homeobox and basic helix-loop-helix (bHLH) domains (Figure 3B; Figure 2). The latter factors are important for regulatory processes during embryogenesis such as neurogenesis, myogenesis, and positional specification along the body axis (Supplementary File 1–Supplementary Table 10). For example, we found the bHLH domain-containing transcription factor MyoD, the master regulator for muscle cell specification in vertebrates, *D. melanogaster*, and *C. elegans* (Tapscott et al., 1988; Michelson et al., 1990; Chen et al., 1994), consistent with the bilaterian origin of mesoderm (Supplementary File 1–Supplementary Table 10, Supplementary File4). Likewise, at least three conserved regulators of nervous system development and neurotransmission, the Neuronal PAS domain-containing protein 4, the Prospero homeobox protein 2, and the Achaete-scute homolog 2 (Stergiopoulos et al., 2014; Sun and Lin, 2016), emerged in the ancestor of bilaterians (Supplementary File 1–Supplementary Table 10, Supplementary File4). Finally, two orthogroups with homeobox domain proteins, OG_8634 and OG_4203, contained the central Hox genes Antennapedia and Ultrabithorax (Balavoine et al., 2002;

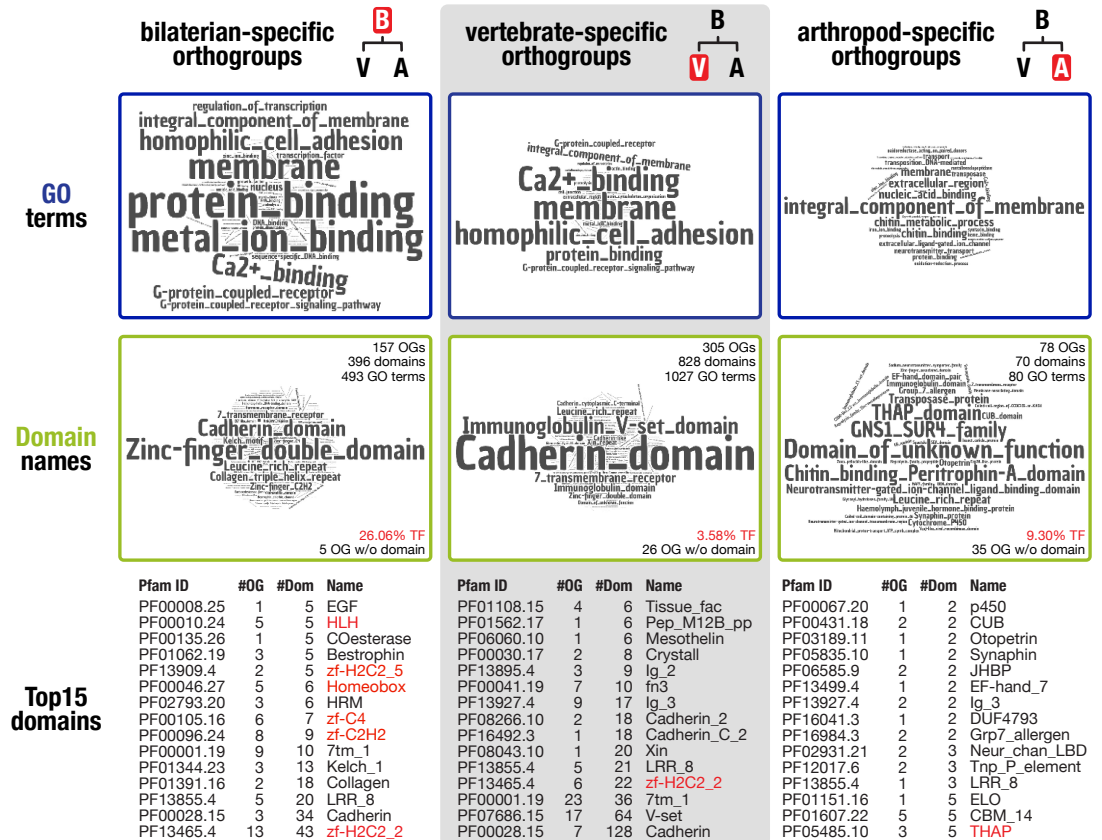


Figure 2. Inventory of protein domains and associated GO terms for three animal lineages. Top row: Relative abundance of GO terms for molecular function for the bilaterian (left), vertebrate (center), and arthropod (right) node. Relationships of the three groups are shown in pictograms on top of each column, with the relevant lineage highlighted in red (B: Bilateria; V: Vertebrata; A: Arthropoda). The GO terms are derived from the domain names (middle row) as determined by domain scans, according to the list <http://geneontology.org/external2go/pfam2go>. Middle row: Relative abundance of domain names found in genes/orthogroups specific for bilaterians (left), vertebrates (center), and arthropods (right). Basic statistics of the respective dataset are shown in the right corners of each panel. #OG w/o domain: number of node-specific orthogroups without known domains (see **Figure 2-Figure Supplement 2**, **Figure 2-Figure Supplement 3**, **Figure 2-Figure Supplement 4**). Bottom row: The 15 most prevalent protein domains of bilaterian-, vertebrate-, and arthropod-specific orthogroups are shown (reverse order). The four columns in each table refer to the official Pfam ID of a domain (Pfam ID), the number of orthogroups with this domain (#OG), the total number of such domains found in all orthogroups specific for a lineage (#Dom), and the common name of the domain (Name). DNA-binding domains as hallmark of transcription factors are highlighted in red. Pep_M12B_pp is short for Pep_M12B_propep (PF01562).

Figure 2-Figure supplement 1. Metazoan poly-zinc finger transcription factor repertoire and evolution.

Figure 2-Figure supplement 2. Multiple sequence alignments of two bilaterian-specific orthogroups without known domains.

Figure 2-Figure supplement 3. Multiple sequence alignments of OG_13336 and OG_31055, two bilaterian-specific orthogroups without known domains.

Figure 2-Figure supplement 4. Multiple sequence alignment of OG_8220, another bilaterian-specific orthogroup without known domains.

Chourrout et al., 2006). Central Hox genes are absent from non-bilaterian Metazoa despite the existence of anterior and posterior homologs (*Ryan et al., 2007*). Our screen thus correctly identified central Hox genes as a bilaterian novelty even though homeodomain-containing proteins are difficult to assign (*Thomas-Chollier et al., 2010; Hueber et al., 2013*).

Membrane factors include neural transducers and novel proteins

A heterogeneous set of proteins was mapped to the membrane compartment (*Figure 3D*). While most of the domains found in 49 orthogroups of this category occurred once or twice, several domains were seen more often, in particular the seven transmembrane receptor domain (7tm; 13x), the leucine-rich repeat (LRR; 5x), the Bestrophin chloride channel (Bestrophin; 3x), and the hormone receptor domain (HRM; 3x). The 7tm domain is characteristic of G protein-coupled receptors, which will be discussed further below. The LRR domain is a protein binding motif (*Kobe and Kajava, 2001*) and present in several factors connected to the plasma membrane (*Figure 3D*) such as LINGO1, SLIT2, or SEMA6C. These LRR domain-containing molecules are crucial for organizing neural connectivity and are employed for axon guidance, myelination, and synapse formation (*de Wit et al., 2011*). Although LRR domain-containing molecules exist in non-bilaterians (*Ocampo et al., 2015*), it is currently unknown whether they fulfil, in these organisms, a role in nervous system development as observed in flies and vertebrates. Further, several bilaterian-specific orthogroups contained ion channel proteins. For both nervous system function and embryonic development (*Moody et al., 1991; Pai et al., 2017*), ion channels play important roles as they provide the basis of currents and action potentials across the plasma membrane and are involved in morphogenetic movements and cell shape changes during development (*Moody et al., 1991*). However, most ion channel proteins seem to predate the origin of metazoans (*Jegla et al., 2009*), and therefore it is unclear how the identified channel proteins affected bilaterian evolution.

Three orthogroups contained transmembrane proteins for which currently no functional description is available, although expression data for two of these exist: OG_13067 (TM169_HUMAN), OG_26661 (TM74B_HUMAN), and OG_28197 (TM160_HUMAN). Genome-wide studies revealed that CG4596, the *Drosophila* ortholog of TM169_HUMAN, is expressed in the ventral nerve cord, ventral midline, and in the brain during embryogenesis (*Tomancak et al., 2002*), similar to central nervous system-based expression of the mouse ortholog (Supplementary File 1–Supplementary Table 11; (*Petryszak et al., 2016*)). Mouse expression data for the transmembrane protein TM160_HUMAN largely overlap with TM169_HUMAN (Supplementary File 1–Supplementary Table 11), but corresponding data from *Drosophila* are not available, as TM160 is absent from ecdysozoans (*Figure 2–Figure Supplement 2*, Supplementary File 1–Supplementary Table 12). Multiple sequence alignments and HMM-HMM searches demonstrate further that these two transmembrane proteins are well conserved across bilaterians (*Figure 2–Figure Supplement 2*) and possess a unique sequence profile without similarity to other orthogroups within the opisthokont search space (Supplementary File 1–Supplementary Table 12). Together, these observations establish that so far uncharacterised proteins with predicted transmembrane domains and distinct structures might have a function in the nervous system since the Cambrian.

Lineage-specific genes are ubiquitous and contain lineage-specific protein domains

The dataset for this study was designed to capture genes with bilaterian-specific distribution. To explore whether it allows the identification of genes specific for other evolutionary nodes, we determined the number of lineage-specific orthogroups for five successive nodes in two lineages: in the protostome lineage leading to Diptera and in the deuterostome lineage leading to Mammalia. We counted for every node lineage-specific orthogroups as a function of increasing species coverage. Extending coverage reduced the number of lineage-specific orthogroups, as expected (*Figure 4*). However, tens to hundreds of lineage-specific orthogroups were still obtained at each individual node under the strict condition of 50 % coverage (i. e. at least 50 % of the species that belong to the respective node need to be present in orthogroups; *Figure 4*). HMM-HMM searches and domain

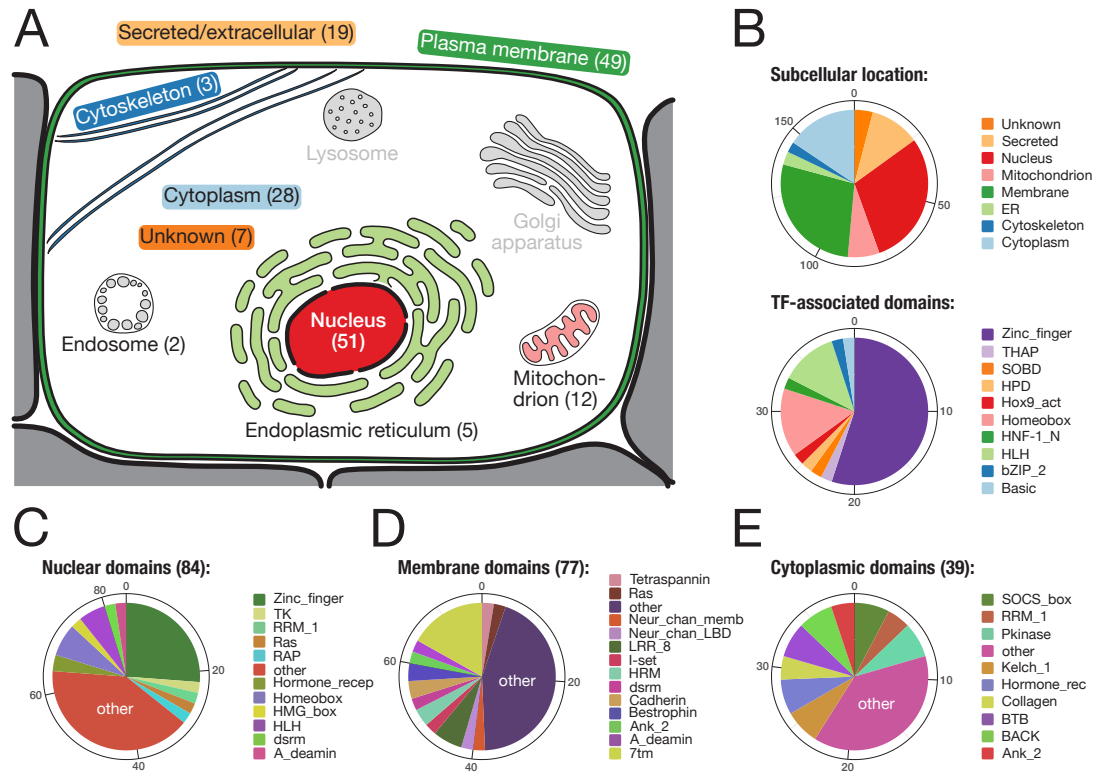


Figure 3. Subcellular location and molecular function of 157 bilaterian-specific genes. **A:** Graphic representation of a eukaryotic cell with its typical organelles. Numbers in parentheses denote the number of bilaterian-specific orthogroups associated predominantly with a given cellular structure. Graphic drawn after the subcellular location section at uniprot.org. **B:** Upper chart: Subcellular location of 157 bilaterian-specific genes. Location data is based on the corresponding human orthologs and colour-matched with the graphics in A. Lower chart: Number and name of transcription factor-associated domains present in the set of 157 bilaterian-specific genes. The 40 orthogroups are a subset of 51 orthogroups associated with the nuclear compartment. In most cases, domains names follow Pfam standards (<http://pfam.xfam.org/>). **C:** Distribution of 84 domains found in 51 orthogroups associated with the nucleus. **D:** Distribution of 77 domains found in 49 orthogroups associated with the plasma membrane. **E:** Distribution of 39 domains found in 28 orthogroups associated with the cytoplasm. «Other» represents domains found only once in the respective category.

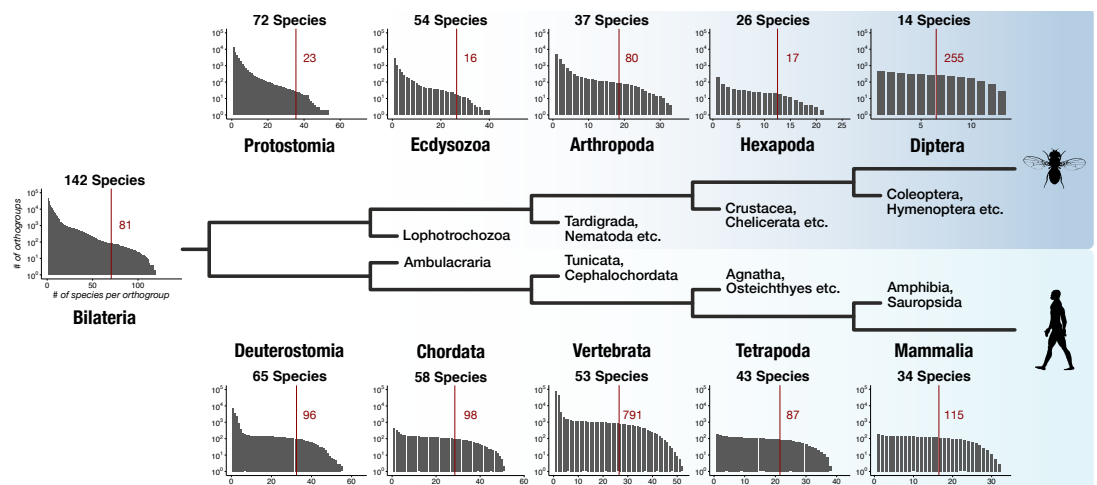


Figure 4. Distinct lineage-specific genes at subsequent nodes of insect and vertebrate evolution.

Starting from Bilateria (left), a protostome lineage leading to dipterans (upper) and a deuterostome lineage leading to mammals (lower) are shown as schematic phylogenetic tree. Sister clades to the selected taxa are denoted on short branches in the center. Each barplot displays the number of lineage-specific orthogroups (y axis) as a function of orthogroup size (x axis) for the selected taxonomic group (Protostomia, Ecdysozoa, Arthropoda etc.). The total species count (within BigWenDB) for each of the eleven taxonomic groups is indicated on top of the corresponding barplots (# Species). The count of lineage-specific genes decreases with growing orthogroup size. A red line denotes the number of orthogroups in which at least 50 % of the species of a selected lineage are present. The corresponding number of lineage-specific orthogroups is highlighted in red next to the line.

Figure 4–Figure supplement 1. Exemplary multiple sequence alignments of three arthropod-specific orthogroups without known domains.

scans further suggested that lineage-specific orthogroups for the ten nodes contain novel domains unique to the respective lineage (for examples, see **Figure 4–Figure Supplement 1** and Supplementary File 1–Supplementary Table 13), as it is the case for bilaterian-specific proteins (**Figure 2–Figure Supplement 2, Figure 2–Figure Supplement 3, Figure 2–Figure Supplement 4**). These findings suggest that the origin of genes and novel protein domains is a robust component of evolution at every examined node and that the faithful identification of these genes is a critical aspect in reconstructing evolutionary history, as exemplified by the recent identification of lineage-specific genes in mammals, mollusks, cnidarians, or arthropods (*Milde et al., 2009; Aguilera et al., 2017; Dunwell et al., 2017; Thomas et al., 2020*).

The Nodal pathway is a bilaterian-specific addition to the TGF- β superfamily and linked to left-right determination and mesoderm formation

Three orthogonal axes—the anterior-posterior, the dorsal-ventral, and the left-right axis—determine body layout in bilaterian animals. One of the signalling systems active in these processes is the Nodal pathway. It belongs to the transforming growth factor β (TGF- β) pathway and is essential for the specification of left-right asymmetry and the induction of mesoderm and endoderm in vertebrates (*Shen, 2007*). The TGF- β ligands Nodal and Lefty, the co-receptor EGF-CFC, and the transcription factor FoxH1 are components specific to the Nodal pathway (**Figure 5–Figure Supplement 1**). In addition, the T-box transcription factor TBR-2/Eomes (T-box brain protein 2/Eomesodermin) is a target of Nodal signalling and critical for mesoderm formation and neural development (*Ryan et al., 1996; Arnold et al., 2008*).

Distinct phylogenetic distributions have been reported for the Nodal-signalling components. The presence and functional conservation of Nodal itself is well established across deuterostomes (*Duboc et al., 2004; Hudson and Yasuo, 2005; Shen, 2007; Röttinger et al., 2015*) and lophotrochozoans (*Grande et al., 2014; Kenny et al., 2014*). In contrast, searches for Lefty orthologs were

so far positive only in deuterostomes (*Chen and Schier, 2002; Mita and Fujiwara, 2007; Duboc et al., 2008; Li et al., 2017*), but not in Lophotrochozoa (*Grande et al., 2014*). Similarly, the Nodal coreceptor EGF-CFC has been identified only in deuterostomes (*Yan et al., 1999; Ravisankar et al., 2011*), and FoxH1 orthologs have been characterized in vertebrates and cephalochordates only (*Weisberg et al., 1998; Zhou et al., 1998; Yu et al., 2008*) (*Figure 5A*). Nodal-signalling components have not been identified in the protostome model organisms *D. melanogaster* and *C. elegans*. Likewise, the T-box factor *eomesodermin* is absent from these animals, but has been described in lophotrochozoans, deuterostomes, and sponges (*Maruyama, 2000; Tagawa et al., 2000; Arenas-Mena, 2008; Arnold et al., 2008; Seb  -Pedr  s et al., 2013*). These findings imply a successive gain of Nodal signalling components along the lineage from the metazoan to the vertebrate ancestor (*Figure 5A*).

In line with previous findings (*Hudson and Yasuo, 2005; Shen, 2007; Grande et al., 2014; Kenny et al., 2014*), our analysis revealed that the TGF- β ligand Nodal belongs to a robust bilaterian-specific orthogroup (OG_12210; *Figure 5–Figure Supplement 2*, Supplementary File 1–Supplementary Table 14). However, orthogroups of the other Nodal pathway members (Lefty, EGF-CFC, FoxH1, and Eomes) were also bilaterian-specific, and HMM-HMM-based searches against all orthogroups (Supplementary File 1–Supplementary Table 14) as well as phylogenetic analyses supported this result (*Figure 5–Figure Supplement 2, Figure 5–Figure Supplement 3*).

Our clustering results suggested further that the T-box transcription factor Eomes is in fact restricted to bilaterians, contradicting a study that identified Eomes candidates in two poriferan species (*Seb  -Pedr  s et al., 2013*). In Blast searches, the two poriferan sequences displayed highest similarity to the canonical T-box transcription factors TBX3/4, but not to the T-box containing protein Eomes (Supplementary File 1–Supplementary Table 15). Likewise, phylogenetic analyses failed to confidently assign the poriferan sequences to the Eomes clade (*Figure 5–Figure Supplement 4*), and HMM-HMM searches could not detect Eomes-related orthogroups with proteins from sponges or other non-bilaterian animals (Supplementary File 1–Supplementary Table 14). These results consistently argue for a bilaterian origin of the factor, matching the distribution of the other Nodal pathway members (*Figure 5B*). While our phylogenetic analyses supported orthology clustering results and the monophyly of the Eomes clade, they unexpectedly argued for a metazoan origin of the gene (*Figure 5–Figure Supplement 4*). This interpretation would imply independent loss events in the ancestors of three phyla (Cnidaria, Placozoa, and Ctenophora) and in two sponge lineages (see *Figure 5A* and discussion), while a posited bilaterian-specific origin would be more parsimonious. To finally resolve this issue, more detailed analyses are needed.

Recently, a Nodal-related gene has been identified in the cnidarian *Hydra magnipapillata* and found to be essential for specifying axial asymmetry along the polyp’s main body axis (*Watanabe et al., 2014*). In our dataset, *H. magnipapillata* Nodal-related belongs to a different orthogroup (OG_9136), together with sequences from nine other cnidarians and many deuterostomes. This orthogroup contains, among others, vertebrate GDF-6/7, but no Nodal orthologs. Furthermore, we did not obtain an HMM-HMM reciprocal best hit relationship with the Nodal orthogroup using as query either the entire orthogroup OG_9136 or a subset of cnidarian sequences (Supplementary File 1–Supplementary Table 16), suggesting that Nodal indeed emerged in the bilaterian ancestor as a new member among pre-existing Nodal-related genes.

Taken together, orthology clustering, HMM-HMM comparison, and phylogenetic evidence establish that all four Nodal-specific pathway components and Eomes are present only in bilaterians (*Figure 5B*). It is thus possible that these factors co-evolved as extension of the more ancient TGF- β signalling pathway (*Huminiecki et al., 2009; Hinck et al., 2016*) and acquired the potential for mesoderm formation and left-right axis determination, two characteristic bilaterian traits. Due to the conservation of this hypothetical gene regulatory network (GRN) since the Cambrian, it could represent an ancient kernel for mesoderm specification and neural patterning. The identification of only a subset of the five factors in non-chordate species (*Figure 5B*) indicates that Nodal signalling experienced substantial evolutionary turnover, but it does not exclude initial assembly of

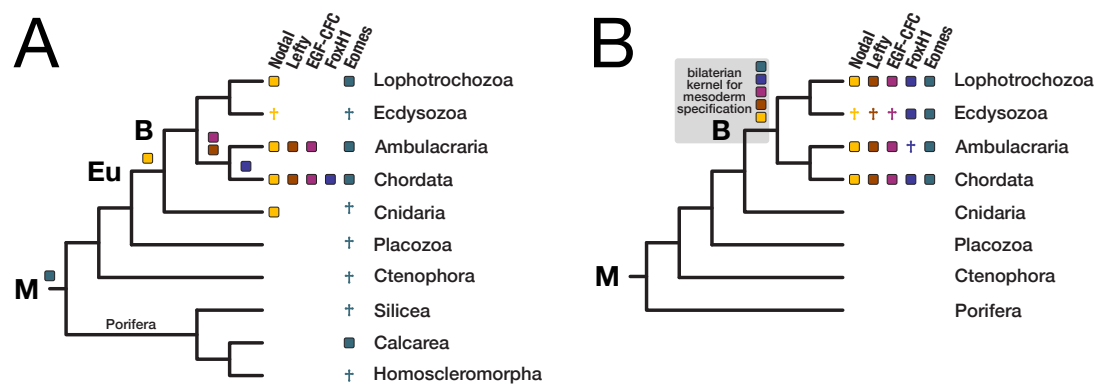


Figure 5. Evolution of the Nodal signalling pathway. Two consensus phylogenetic trees showing the relationship of major metazoan lineages. The five factors of the Nodal signalling pathway (Nodal, Lefty, EGF-CFC, FoxH1, and Eomes) are displayed as coloured boxes. Their phylogenetic distribution and inferred evolutionary origin are mapped onto the tree. Gene births are indicated as coloured boxes above the respective branch. Inferred losses are represented by crosses. Bold labels to the left of a branch indicate branch ancestors: B = Bilateria, Eu = Eumetazoa, M = Metazoa. **A:** Previous results regarding the evolution of Nodal pathway genes, as known from the literature. **B:** Revised evolutionary history of the Nodal pathway genes according to our results. Note that none of the five factors has been found in arthropods and nematodes. The ecdysozoan boxes for Eomes and FoxH1 are derived from the presence of the genes in a single priapulid species. Grey shading: Hypothetical emergence of a putative kernel for mesoderm specification and neural patterning.

Figure 5-Figure supplement 1. Schematic outline of the Nodal signalling pathway in vertebrates.

Figure 5-Figure supplement 2. Bilaterian-specific distribution of the Nodal pathway components Nodal and Lefty.

Figure 5-Figure supplement 3. Bilaterian-specific distribution of the Nodal pathway component FoxH1.

Figure 5-Figure supplement 4. Bilaterian-specific distribution of the Nodal pathway component Eomesodermin.

the pathway in the bilaterian ancestor and subsequent lineage-specific changes.

One consequence of these considerations is that large parts of the Nodal GRN must have been lost early in ecdysozoan evolution, implying the evolution of alternative upstream signalling pathway inputs for axial specification in this group. Secondly, genes that originated in the bilaterian ancestor may have been lost in a particular daughter lineage. The widespread loss of genes across metazoans (Richter et al., 2018; Sharma et al., 2018) and the loss of Nodal pathway members [this study] shows that such scenarios are conceivable and might impact the exhaustive description of lineage-specific genes, i. e. the reconstruction of the «true» evolutionary history of a taxon.

G protein-coupled receptors and the control of physiological state through circulatory flow

Among the identified bilaterian-specific genes is a set of eight G protein-coupled receptors (GPCRs), members of a large family of seven-transmembrane domain receptors. While GPCRs are ancient and were already present in the ancestor of bilaterians and fungi (Krishnan et al., 2012), our results indicate that new members of the GPCR family appeared at the bilaterian base. Specifically, robust clustering results and HMM-HMM comparisons place the origin of monoamine neurotransmitter receptors for serotonin, adrenaline, and dopamine to the bilaterian root (Supplementary File 1–Supplementary Table 17, Supplementary File 1–Supplementary Table 18), in line with a recent publication that dated back the evolutionary history of adrenergic signalling to the bilaterian ancestor (Bauknecht and Jékely, 2017). Histochemical, biochemical, and functional data are in conflict with this finding and argue for the presence of serotonin, dopamine, and other small molecule neurotransmitters in cnidarians, the bilaterian sister group (Carlberg and Anttil, 1993; Kass-Simon and Pierobon, 2007; Mayorova and Kosevich, 2013). However, receptors for these molecules could

not be identified unambiguously in cnidarians (*Anctil, 2009; Bosch et al., 2017*), maintaining the possibility that they indeed constitute bilaterian innovations.

There is evidence across several bilaterian phyla (arthropods, nematodes, mollusks, platyhelminthes, vertebrates) that adrenaline, dopamine, and serotonin signalling regulates many important processes such as behaviour, feeding, learning, locomotion, memory, reproduction, reward, or sleep (*Ségalat et al., 1995; Berridge, 2004; Suo et al., 2004; Berger et al., 2009; Vidal-Gadea et al., 2011; Burke et al., 2012; El-Shehabi et al., 2012; Ueno et al., 2012*). In addition to these «post-embryonic» functions, serotonin is recognised as an important regulator of embryonic development and neuronal circuitry in vertebrates and invertebrates (*Brown and Shaver, 1989; Buznikov et al., 2001; Daubert and Condron, 2010*). The proposed origin of monoamine neurotransmitter receptors in the bilaterian ancestor (Supplementary File 1–Supplementary Table 17, Supplementary File 1–Supplementary Table 18) and the related functions of monoamine neurotransmitter signalling across phyla suggest that diverse functions of monoamine neurotransmitter signalling already existed in the bilaterian ancestor and could have played a role in the evolution of complex development, brain function, and behaviour. Preliminary evidence indicates that cnidarians, as the bilaterian sister group, do not respond to rewarding or punishing stimuli as do bilaterians (*Barron et al., 2010*). A link between this behavioural difference and the evolution of monoamine neurotransmitter receptors would comply with the previous notion that the evolution of dopamine-based brain reward systems in bilaterians started from dopamine's ancient role as a signalling molecule for motor circuits (*Barron et al., 2010*).

In addition to monoamine neurotransmitter receptors, we detected several peptide hormone receptors in the set of bilaterian-specific GPCRs and underscored their bilaterian origin using HMM-HMM searches: the receptors for secretin, corticotropin-releasing factor, neuromedin-U, calcitonin, and somatostatin (Supplementary File 4, Supplementary File 1–Supplementary Table 17, Supplementary File 1–Supplementary Table 18). In vertebrates, these GPCRs and their hormone ligands are part of the endocrine system and regulate basal physiological activities such as feeding, energy homeostasis, or stress (*Budhiraja and Chugh, 2009; Afroze et al., 2013*). Homologs of the five receptors and their ligands have also been described in *C. elegans* and *D. melanogaster* (*Johnson et al., 2005; Cardoso et al., 2006; Melcher et al., 2006; Lindemans et al., 2009; Cardoso et al., 2014; Kunst et al., 2014; Ketchesin et al., 2017*), and the putative bilaterian ancestry of some of these signalling systems has been recognised by others, in agreement with our results (*Johnson et al., 2005; Lindemans et al., 2009; Mirabeau and Joly, 2013*). In contrast to vertebrates or insects, cnidarians and other non-bilaterian Metazoa do not contain specialized endocrine organs and circulatory systems. Thus, our finding of highly conserved peptide hormone receptors supports the view that major physiological regulators evolved in parallel with the emergence of circulatory systems. Moreover, recent evidence indicates that these hormone receptors also act during development and participate in neuronal migration and nervous system formation (*Afroze et al., 2013; Liguz-Leczna et al., 2016; Galas et al., 2017*), suggesting an ancient link between the generation of complex nervous systems and the ability to control body functions through circulatory fluid.

Changes in axon guidance accompany bilaterian evolution

Axon guidance, the guided outgrowth of axons and dendrites, is essential for the development of neuronal connections and mediated by two major pathways, the Netrin-DCC and the Slit-Robo (Round-About) pathway (*Lowery and Van Vactor, 2009; Evans, 2016*). To reveal whether changes in these processes accompanied the evolution of bilaterians, we studied the respective orthogroups. Except one, all human Netrin paralogs were assigned to a single orthogroup. Its composition and the composition of its HMM-HMM best hit orthogroups support the emergence of Netrins in the ancestor of eumetazoans or earlier (Supplementary File 1–Supplementary Table 19), in line with a description of Netrins in the sea anemone *N. vectensis* (*Putnam et al., 2007*). We found a corresponding (eu)metazoan origin for the Netrin receptor DCC (Supplementary File 1–Supplementary Table 19). These results indicate that cnidarians, but not ctenophores, might regulate axon out-

growth at least in part by Netrin-DCC based interactions, consistent with an independent origin of the nervous system in ctenophores (*Moroz et al., 2014*).

Although orthogroup composition of Slit and its receptor Robo suggested a bilaterian origin of this system, reciprocal HMM-HMM searches indicated the existence of cnidarian Robo orthologs that were assigned to a separate orthogroup, OG_51853 (Supplementary File 1–Supplementary Table 19). Like their bilaterian counterparts, the cnidarian Robo candidates had highly disordered cytoplasmic domains, as revealed by structure predictions of the extracellular and intracellular parts of representative sequences (*Figure 6*). On the other hand, sequence comparisons revealed that the conserved cytoplasmic motif CC1, which is required for binding the Ena/VASP protein Enabled and for transducing signals to the actin cytoskeleton (*Bashaw et al., 2000*), is altered in cnidarian Robos (*Figure 6–Figure Supplement 1*), and that cnidarian Robos displayed several insertions and deletions in the cytoplasmic part when compared with bilaterian Robos (*Figure 6–Figure Supplement 2*). It is therefore an open question whether the structural differences in cnidarian Robo-like proteins involve interactions with different downstream partners and whether cnidarian Robos regulate axon growth. Known downstream effectors of Robo signalling, such as Enabled and Son of sevenless, originated early in metazoan evolution (Supplementary File 1–Supplementary Table 20) and could provide in principle the functionality for Robo-based axon guidance, although mediated by a different ligand.

In both *D. melanogaster* and vertebrates, midline glia cells secrete the Slit protein to prevent Robo expressing axons from crossing the body midline (*Rothberg et al., 1990; Brose et al., 1999; Kidd et al., 1999*), indicating that a key component in the establishment of bilaterally symmetric nervous systems is shared between protostomes and deuterostomes. However, in our dataset a single placozoan sequence was assigned to Slit's otherwise bilaterian-specific orthogroup, shifting its origin back in time. Blast searches at NCBI verified a reciprocal best hit relationship of the putative placozoan Slit to known Slit proteins, in agreement with our clustering results (Supplementary File 1–Supplementary Table 15). Likewise, placement of the placozoan sequence in phylogenetic analyses is compatible with its orthology to the Slit protein (*Figure 6–Figure Supplement 3*). Unexpectedly, HMM-HMM comparisons could not reveal the existence of Slit in other non-bilaterian species such as cnidarians or ctenophores (Supplementary File 1–Supplementary Table 21). From these results we conclude that Slit and Robo probably originated in the common ancestor of placozoans, cnidarians, and bilaterians. However, the Slit-Robo-based mechanism for midline repulsion during nervous system development appears to be restricted to bilaterians, as placozoans lack a nervous system and cnidarians lack the Slit ligand.

Neurotrophin receptor signalling is a bilaterian innovation

Neurotrophin signalling plays a fundamental role in nervous system generation by regulating many aspects of neuronal development and function, such as neuronal survival, synapse formation, or axon guidance (*Huang and Reichardt, 2001; Lu et al., 2005*). Vertebrates possess four related neurotrophin ligands and three corresponding transmembrane receptors of the Trk family that each originated from a single ancestral gene in chordates (*Benito-Gutiérrez et al., 2005; Hallböök et al., 2006*). Once considered a vertebrate innovation, neurotrophins and their receptors have now been found in diverse invertebrates (*Wilson, 2009; Kassabov et al., 2013; Lauri et al., 2016*). In particular, studies in the mollusk *Aplysia californica* suggest that neurotrophin signalling and neurotrophin-mediated synaptic plasticity are conserved in protostomes and deuterostomes (*Kassabov et al., 2013*).

To elucidate the evolutionary origin of neurotrophin signalling, we analysed the orthogroups containing neurotrophins and their receptors. The four vertebrate neurotrophin ligands clustered into two bilaterian-specific orthogroups (OG_14798 and OG_21801) that are each other's reciprocal best hit. We could not detect orthogroups similar to neurotrophins in non-bilaterian metazoans or additional, so far unidentified neurotrophins in bilaterians (Supplementary File 1–Supplementary Table 22), supporting the emergence of a single neurotrophin gene in the ancestor of bilateri-

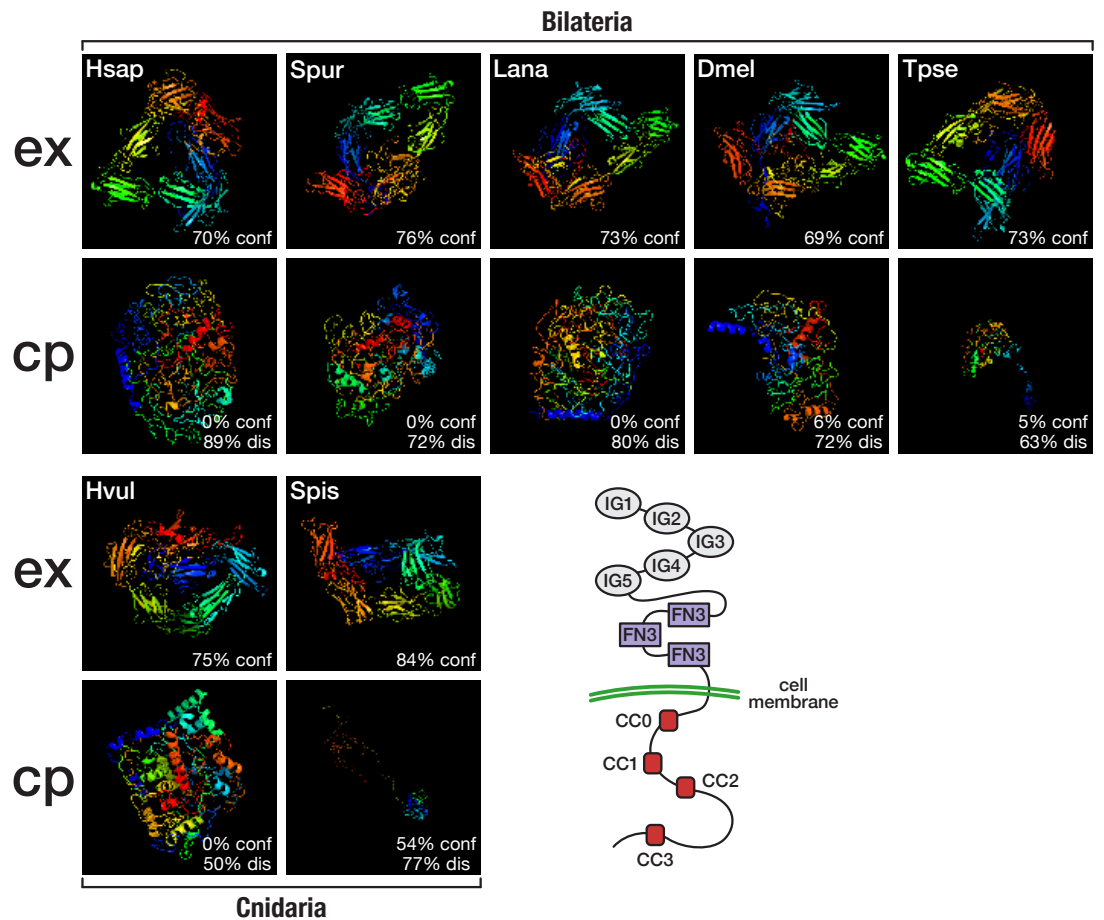


Figure 6. Structural predictions of cnidarian and bilaterian Robo proteins. Top (ex): Predicted structure of the extracellular domain plus transmembrane region of seven selected Robo proteins. Bottom (cp): Predicted structure of the transmembrane region plus cytoplasmic part of seven selected Robo proteins. Robo1 orthologs of two deuterostomes (Hsap = *Homo sapiens*; Spur = *Strongylocentrotus purpuratus*), one lophotrochozoan (Lana = *Lingula anatina*), two ecdysozoans (Dmel = *Drosophila melanogaster*; Tpse = *Trichinella pseudospiralis*), and two cnidarians (Hvul = *Hydra vulgaris*; Spis = *Stylophora pistillata*) were analysed. «% conf» indicates the percentage of residues modelled at >90 % confidence. «% dis» indicates the predicted percentage of disordered regions. Bottom right: Schematic outline of the Robo domain structure with five immunoglobulin domains (IG1–IG5) and three fibronectin type III domains (FN3) in the extracellular part and four conserved cytoplasmic motifs (CC0–CC3) in the intracellular part. Like their bilaterian counterparts, cnidarian Robo candidates display disorganized protein structure in the cytoplasmic part despite differences in structural features (Figure 6–Figure Supplement 1, Figure 6–Figure Supplement 2). The extracellular part (top row), on the other hand, is similarly organized across metazoans.

Figure 6–Figure supplement 1. Change of the conserved cytoplasmic motif CC1 in cnidarian Robo-like proteins.

Figure 6–Figure supplement 2. Cnidarian Robo-like proteins display structural alterations.

Figure 6–Figure supplement 3. Phylogenetic analysis of a putative *Trichoplax adhaerens* Slit protein.

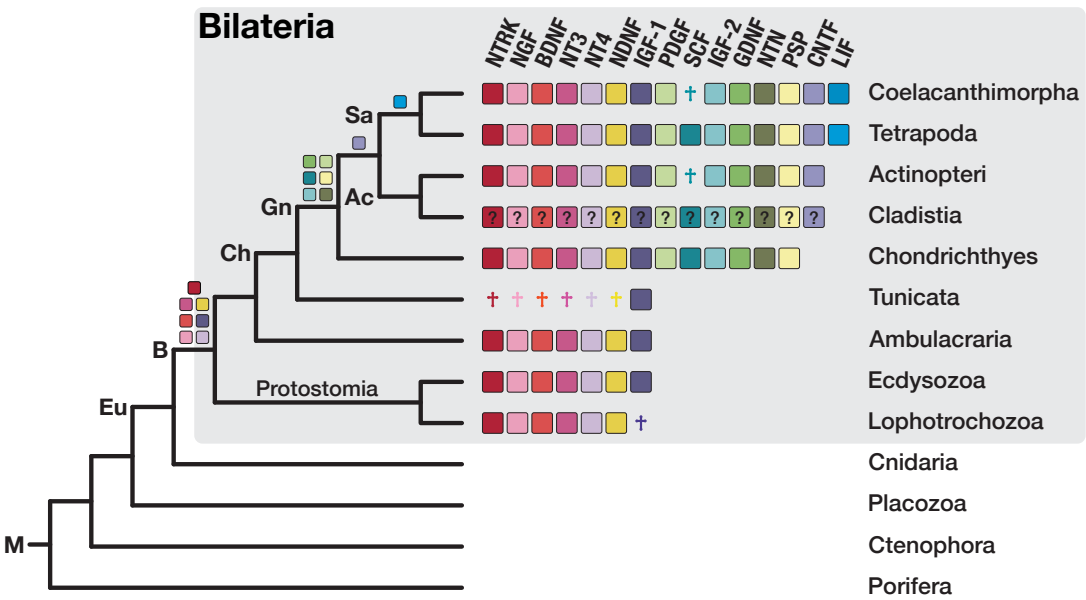


Figure 7. The bilaterian-wide distribution of neurotrophic factors. The NTRK receptor and 14 major neurotrophic factors are displayed as coloured boxes. Their phylogenetic distribution and inferred evolutionary origin are mapped onto the tree (see Supplementary File 1–Supplementary Table 22 and Supplementary File 1–Supplementary Table 23). Gene births are indicated as coloured boxes above the respective branch of the tree (left). Inferred losses are shown as coloured crosses in the matrix. Bold labels to the left of a branch indicate branch ancestors: Ac = Actinopterygii; B = Bilateria; Ch = Chordata; Eu = Eumetazoa; Gn = Gnathostomata; M = Metazoa; Sa = Sarcophagaria. The neurotrophic factors of Cladistia, the sister group of Actinopteri, are inferred and distinguished by a question mark as the dataset lacks species from this lineage.

Figure 7–Figure supplement 1. The NTRK neurotrophin receptor is restricted to bilaterians.

ans and its subsequent diversification in vertebrates. When we analysed the evolutionary origin of other neurotrophic factors, we recognised that they also arose in the ancestor of bilaterians or even later (Figure 7; Supplementary File 1–Supplementary Table 22, Supplementary File 1–Supplementary Table 23). The evolutionary age of these additional neurotrophic factors is thus consistent with a bilaterian origin of neurotrophic ligands per se. The same evolutionary scenario is supported by detailed analysis of the Trk receptor family. Although our initial dataset conflated Trk and Wnt pathway receptors due to a shared receptor tyrosine kinase domain, adjustment of the MCL inflation parameter successfully rendered a Trk-only orthogroup, whose taxonomic composition is restricted to bilaterians (Figure 7–Figure Supplement 1; Supplementary File 1–Supplementary Table 24).

These results indicate that neurotrophins and their receptors are present across bilaterians and might fulfill conserved functions in neuronal development in these animals. If long-term potentiation and memory formation is regulated by serotonin and its receptors across bilaterians (see, for example, Teixeira et al., 2018), a link between serotonin action and neurotrophin signalling may have emerged in the bilaterian ancestor that contributed to nervous system evolution and the learning-dependent synaptic plasticity characteristic for this group.

Bilaterian-specific factors and the evolution of excretory systems

Protostomes and deuterostomes comprise the taxon Nephrozoa, animals with a dedicated excretory system (sensu Jondelius et al., 2002). Together with their sister group Xenacoelomorpha, Nephrozoa form the taxon Bilateria (Cannon et al., 2016). When we started with our study, sequences from Xenacoelomorpha were not available, and therefore our bilaterian-specific gene set is in fact specific for nephrozoans and might contain factors related to kidney and/or nephron

development. Indeed, we identified in the 157 bilaterian-specific orthogroups two relevant zinc finger transcription factors. The poly-zinc finger transcription factor Evi1/MECOM was assigned to a large orthogroup with protein members from 108 of 142 bilaterian species (OG_5543). Evi1 is expressed in pronephric tissue of *Xenopus* and zebrafish embryos and involved in nephron patterning in these species (Mead et al., 2005; Li et al., 2014; Desgrange and Cereghini, 2015), although this might only be a part of its function (Goyama et al., 2008). Secondly, after Blast searches, maximum likelihood phylogenetic analysis, and HMM-HMM searches focusing on orthogroup OG_5226, we found evidence for a bilaterian-wide distribution of odd-skipped related 1, a zinc finger transcription factor required for heart and urogenital development in vertebrates (Wang et al., 2005; Dressler, 2006; Tena et al., 2007) (Supplementary File 1–Supplementary Table 15, Supplementary File 1–Supplementary Table 26; Supplementary File 1–Supplementary Figure 1). Thus, the observed expansion of the zinc finger transcription factor repertoire may also have been important for the evolution and development of excretory organs, a key nephrozoan innovation.

Bilaterian-specific genes form a rich interaction network with interconnected sub-networks

To analyse potential interactions among the 157 bilaterian-specific proteins, we obtained the corresponding human orthologs and analysed their interaction network using the STRING protein-protein interaction (PPI) database. The obtained PPI network contained significantly more interactions than expected by chance (PPI enrichment p-value: $5.93e^{-14}$), revealing that bilaterian-specific genes form a dense network in which about 50 % of the factors (83 distinct factors) are connected to one another (Figure 8A). These interactions form several subnetworks involved in regulating key aspects of bilaterian development, such as chromatin organization and transcriptional regulation (subnetwork A), myogenesis (subnetwork B), mesoderm formation and left-right asymmetry (the Nodal pathway, subnetwork C: see also Figure 8B), neurogenesis (subnetwork D), and physiology (subnetwork E). Connections between different subnetworks further suggest that crosstalk between the newly established regulatory subnetworks was an important aspect of bilaterian evolution.

Previous work found that protein network connectivity (number of interactions) increases with gene age (Kim and Marcotte, 2008). To analyse the degree of connectivity of our bilaterian network, we compared it to a PPI network generated from metazoan-specific proteins that is expected to show higher connectivity due to the proteins' more ancient origin. Our orthology clustering data identified 797 metazoan-specific proteins (>5× as many proteins as in the bilaterian dataset), and the combined bilaterian-metazoan PPI network comprised 2,531 interactions among 823 proteins (16 % bilaterian-specific proteins, 84 % metazoan-specific proteins). In fact, we obtained a slightly higher level of connectivity for the younger, bilaterian proteins (Figure 8C: total number of interactions per protein, median ± median absolute deviation (MAD): 5 ± 4.62 for Bilateria, 4 ± 4.16 for Metazoa; Mann-Whitney U test: $U = 39792$, $p = 0.0135$). Furthermore, bilaterian-specific proteins preferentially interacted with one another, with over twice as many bilaterian-bilaterian interactions as would be expected by chance (χ^2 statistic = 24.814, $p = 0.000001$), primarily due to fewer bilaterian-metazoan interactions than would be expected. This is also evident at the level of individual proteins: bilaterian-specific proteins have significantly more bilaterian interaction partners (Figure 8D: percent of bilaterian interactions, median ± MAD: $19.5\% \pm 23.2\%$ for Bilateria, $0.0\% \pm 16.1\%$ for Metazoa; Mann-Whitney $U = 32231$, $p = 0.00000$).

As we identify the Nodal pathway as a key bilaterian innovation (Figure 5, Figure 8A: subnetwork C), we focused on this subnetwork as a case study for further analysis of molecular interactions. Within the full bilaterian-metazoan PPI network, we indeed recovered the Nodal pathway as a bilaterian-specific subnetwork, embedded among connections to additional bilaterian and metazoan proteins (Figure 8B). As with the full network, for this subnetwork we found a significant number of bilaterian-specific protein interactions (Figure 8D; Kruskal-Wallis $\chi^2 = 62.855$, degrees of freedom = 3, $p = 1.44e^{-13}$). Furthermore, for this subnetwork, we found support for the hypothesis that

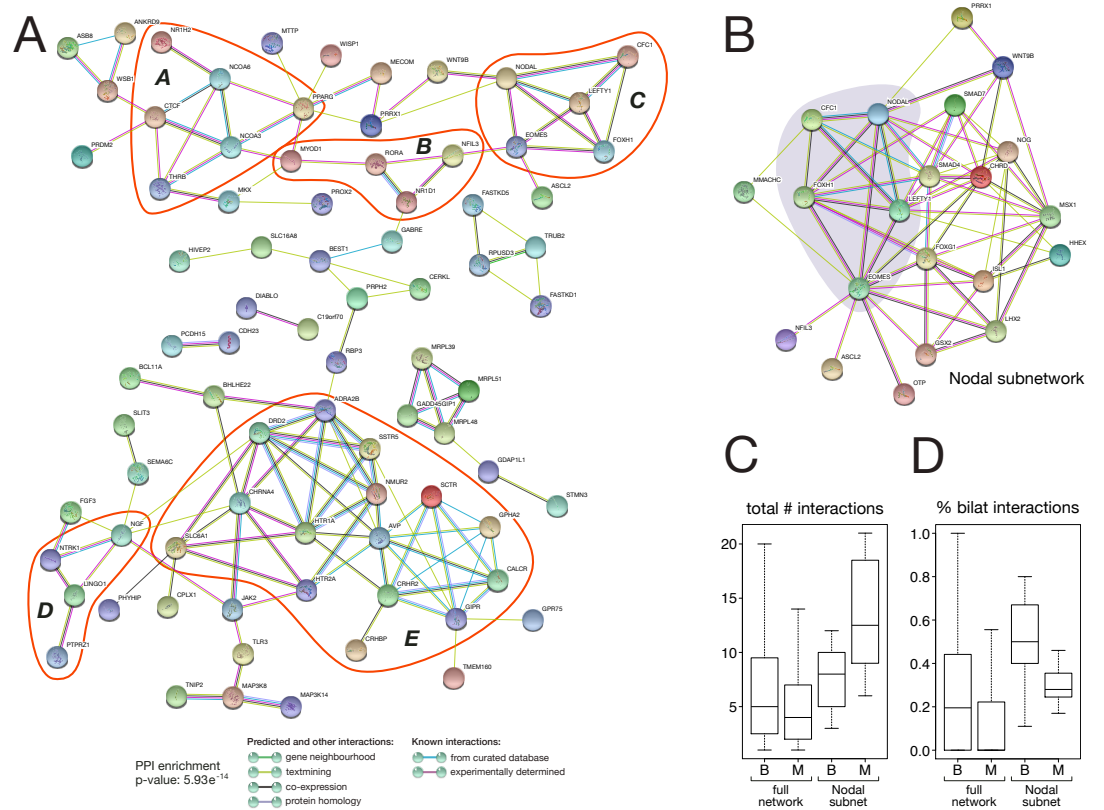


Figure 8. Protein-protein interaction network of bilaterian-specific proteins. **A:** Uniprot identifiers corresponding to the human orthologs of 150 bilaterian-specific genes (seven OGs had no human ortholog) were uploaded to the STRING database, and their mutual interactions were visualized as a network. Parameters for the displayed PPI network were: minimum required interaction score = 0.4; maximum number of interactors to display in 1st and 2nd shell = 0. Thus, only known and predicted interactions between 83 distinct bilaterian-specific proteins are shown (non-interacting proteins are hidden). Evidence for displayed interactions is colour-coded (see legend). Edge length and node placement are arbitrary. Five subnetworks between bilaterian-specific genes are highlighted in red (A-E, see Results). **B:** Bilaterian-specific Nodal subnetwork in the context of metazoan genes. The five members of the Nodal pathway are highlighted by shading. **C, D:** Boxplots comparing bilaterian- (B) and metazoan-specific (M) proteins in the full network and Nodal subnetwork for the total number of interactions per protein (C), and for the relative fraction of bilaterian interactions per protein (D).

older (metazoan) genes have higher connectivity (*Figure 8C*; *Kim and Marcotte (2008)*). Notably, metazoan-specific proteins that participate in the Nodal subnetwork are a non-representative subset, showing significantly higher overall connectivity and bilaterian-specific connectivity than metazoan proteins in the full bilaterian-metazoan PPI network. Thus, it may be that older genes have higher connectivity if they exceed a minimum threshold of connectivity (number of interactions). For example, the Nodal subnetwork includes Smad4, a metazoan-specific protein with the highest connectivity (46 interactions) of any protein in our combined network. This multifunctional BMP pathway component likely exemplifies two evolutionary trends: that highly connected genes are most likely to acquire new interaction partners, and that bilaterian-specific PPI innovations build on more ancient, preexisting PPI networks by co-option.

Extrapolating these findings to interactions with additional factors of more ancient origin implies that the evolution of new genes in the bilaterian ancestor affected a large number of processes in animal biology.

Discussion

An R-based OrthoMCL pipeline for processing large datasets

Explaining the sudden emergence of bilaterally symmetric animals during the Cambrian is a central problem in evolutionary biology. Complicated by the uneven coverage of the metazoan tree with sequence information, a systematic approach to identify the genetic basis for the evolution of bilaterians was missing. In this study, we present a comparative genomics approach, designed to provide maximum resolution at the bilaterian/non-bilaterian divergence and therefore uniquely suited to discover bilaterian-specific genes.

Although sequence data for individual species in our study might be incomplete (Supplementary File 1–Supplementary Table 1, Supplementary File 1–Supplementary Table 2), each important taxonomic group (Deuterostomia, Ecdysozoa, Lophotrochozoa, and «non-Bilateria») is represented with several well-annotated genomes and/or proteomes (**Figure 1–Figure Supplement 1**, Supplementary File 1–Supplementary Table 3). Importantly, sequence data from 19 cnidarian species, including four sequenced genomes and five transcriptomes with CEGMA scores above 70 % (Supplementary File 1–Supplementary Table 2), allow the crucial distinction of orthogroups with cnidarian participation from bilaterian-specific orthogroups without cnidarian contribution, a serious problem of existing databases (**Table 1**).

While other orthology databases might surpass the BigWenDB in species number, this is often due to the integration of many non-metazoan and prokaryotic species (**Table 1**). Still, the total sequence content of other databases is small enough to be handled by a MySQL engine (see <http://www.orthodb.org/v9.1/download/README.MySQL.txt>; www.orthomcl.org) because it is restricted to predicted and annotated protein sequences. To accomplish processing of the large amount of sequence data from 25 genomic ORF sets, we developed an R-based version of the OrthoMCL pipeline (**Li et al., 2003**). It reproduces the results of the original pipeline meticulously (Supplementary File 1–Supplementary Table 4) and is capable of processing at least 125 million sequences with current computer hardware, considerably extending the limit imposed by conventional MySQL usage. In view of the ongoing increase in sequence data, the R-based version of OrthoMCL may prove valuable for generating large and comprehensive orthology datasets in the future.

Importantly, scaling up the orthology engine to handle larger datasets did not come at the expense of clustering quality. Rather, the combination of a comprehensive dataset and a scalable orthology prediction tool turned out as beneficial, challenging an early study that found a high false-positive rate when testing OrthoMCL on a small and taxonomically restricted dataset (**Chen et al., 2007**). This advance is further demonstrated by correct orthology inference rates with our approach that surpass those previously obtained in the orthobench comparisons (**Trachana et al. (2011)**; Supplementary File 3).

Reciprocal HMM-HMM comparisons for improving orthogroup completeness

Despite the existence of many orthology detection methods (**Tekaia, 2016**), current tools do not evaluate orthogroup composition after clustering. In contrast, we implemented filtering steps to first identify widely distributed bilaterian-specific orthogroups. We then applied to the resulting orthogroups extensive procedures for quality control and error correction, taking into account the taxonomic composition of orthogroups and their best hits in HMM-HMM searches. In this context, we developed a new reciprocal HMM-HMM comparison step to evaluate orthogroup completeness because reliable orthogroups are a prerequisite for inferring the evolutionary age of the corresponding gene (Supplementary File 1–Supplementary Table 7). Although HMMs generated from orthogroup alignments can be uninformative outside conserved regions, they capture important amino acid positions and their spacing and variability, and therefore the individual profile of an orthogroup even within common functional domains such as zinc fingers (Supplementary File 1–Supplementary Figure 2). Indeed, we observed several instances where HMM-HMM

comparisons improved results and affected conclusions, demonstrating the value of this novel step (Supplementary File 1–Supplementary Table 13, Supplementary File 1–Supplementary Table 14, Supplementary File 1–Supplementary Table 16, Supplementary File 1–Supplementary Table 19, Supplementary File 1–Supplementary Table 21, Supplementary File 1–Supplementary Table 22, Supplementary File 1–Supplementary Table 23, Supplementary File 1–Supplementary Table 24).

In particular, we employed highly sensitive HMM-HMM comparisons to minimize errors caused by low protein traceability, the limitation of the Blast algorithm to detect orthologous genes in distantly related organisms (Jain et al., 2019; Weisman et al., 2020). This strategy led to the removal of 68 false-positive orthogroups from an initial set of 431 bilaterian-specific orthogroups because they displayed reciprocal best-hit relationships to non-bilaterian orthogroups, indicating a more ancient origin (see Appendix 1: Identification of bilaterian-specific genes). In addition, the broad coverage of bilaterians and non-bilaterians and the evaluation of orthogroup composition by filtering rules minimizes errors that may be caused by the low traceability of specific genes or by single taxa with particularly high evolutionary rates.

Limitations of our orthology clustering pipeline

Our methods for error correction facilitate the detection of reliable lineage-specific gene sets and may serve as a future standard. However, developing software that can automatically detect such patterns and combine/split orthogroups in awareness of the underlying phylogeny would further improve orthogroup assignments. That lineage-specific genes exist and can directly change an animal's phenotype to gain access to new ecological niches has been shown recently, illustrating the importance of these genes and the need for their identification (Dunwell et al., 2017; Santos et al., 2017; Luis Villanueva-Cañas et al., 2017).

Although we obtained a robust set of 157 genes that evolved in the bilaterian ancestor or, more specifically, in the ancestor of protostomes and deuterostomes (Nephrozoa) (Jondelius et al., 2002), by design our study is limited to protein coding sequences. It will therefore miss the possible involvement of RNA genes in bilaterian evolution, including miRNAs (micro RNAs) and lncRNAs (long non-coding RNAs), as suggested by Prochnik et al. (2007). It will further fail to detect changes in cis-regulatory regions and structural alterations or epigenetic changes, additional factors that affect evolutionary processes (Carroll, 1995; Prud'homme et al., 2006; Klironomos et al., 2013; Feulner and De-Kayne, 2017). Despite these limitations, our study successfully corroborated the bilaterian origin of several previously known bilaterian-specific genes, such as the chromatin organizer CTCF (Heger et al., 2012), the left-right determination factor Nodal (Grande et al., 2014), or central Hox genes (Finnerty and Martindale, 1999; Hueber et al., 2013).

Challenges in reconciling orthogroups and phylogenetic trees

Orthology clustering is a distinct method from phylogenetic tree building, and when we used phylogenetic analyses to validate orthogroup composition, we experienced difficulties in reconciling the two approaches.

Firstly, we do consistently obtain high branch support for bilaterian-specific orthogroups as discrete clades. Yet within orthogroups, phylogenetic resolution was often weak, with low branch support and gene tree–species tree discordance. However, tree discordance in itself does not argue against orthology because phylogenies suffer from various problems, such as the inclusion of problematic sequences, little phylogenetic information, or—in our case—the presence of short ORF fragments (Aguileta et al., 2008; Som, 2014). While our ORF data help the recognition of distinct orthogroups by avoiding systemic annotation errors from external databases and by providing essential taxonomic coverage, these sequences do not represent full-length proteins and may curtail within-orthogroup resolving power.

In addition, in several cases we obtained tree topologies that could imply orthogroup origin in the metazoan ancestor rather than a later, bilaterian origin (Figure 5–Figure Supplement 3, Figure 5–Figure Supplement 4, Figure 7–Figure Supplement 1). One major confounding factor for

correct tree reconstruction is heterotachy: a non-constant rate of evolution among different lineages (Lopez *et al.*, 2002; Wu and Susko, 2011; Jayaswal *et al.*, 2014). Importantly, heterotachy is often observed along the branches originating from a gene duplication event (Kondrashov *et al.*, 2002; Conant and Wagner, 2003; He and Zhang, 2005; Steinke *et al.*, 2006). Accelerated evolution in bilaterian-specific «duplicates» could therefore explain the observed tree topologies and the discrepancy between trees and clustering results. In contrast, the alternative interpretation of metazoan orthogroup origins would require that one of the two duplicates was secondarily lost in the stem lineage of sponges, ctenophores, placozoans, and cnidarians because of its absence in all available samples from these phyla. Gene loss is increasingly recognized as a widespread and important evolutionary mechanism (Sharma *et al.*, 2018; Hecker *et al.*, 2019; Thomas *et al.*, 2020). However, the loss of a number of genes in the stem lineages of four independent phyla would imply strong selective pressure against their presence in non-bilaterian lineages, creating an aspect of deep evolution worthwhile of future exploration.

A robust association between bilaterian-specific genes and key morphological features

Several morphological features are widely considered key bilaterian innovations: (i) a third germ layer, the mesoderm; (ii) a complex bilateral nervous system; (iii) a Hox gene cluster with at least seven anterior, posterior, and central Hox genes; (iv) a through gut; (v) an excretory system; (vi) the possession of many different cell types; and (vii) bilateral symmetry (Baguña *et al.*, 2008, and references therein). It was unknown so far whether, and if so which, genetic factors contributed to the emergence of these innovations. From the results presented here, we conclude that a considerable fraction of the identified 157 bilaterian-specific genes is associated with the origin of characteristic bilaterian traits. Although correlations cannot prove a causal relationship, in the absence of ancestral genetic information our inferences from extant animals offer a fruitful approach. Here, we elaborate on several instances where the origin of proteins and bilaterian traits appear to coincide.

For example, a large portion of the 157 genes is involved in nervous system development and/or maintenance (Supplementary File 4). Several factors in this category provide functionalities absent from non-bilaterian metazoans, such as the long-range control of behaviour and physiological state through an expanded repertoire of GPCRs (Supplementary File 1–Supplementary Table 17, Supplementary File 1–Supplementary Table 18), a midline repulsion mechanism for the establishment of a bilateral nervous system (Robo-Slit; *Figure 6–Figure Supplement 3*; Supplementary File 1–Supplementary Table 19, Supplementary File 1–Supplementary Table 21), or mechanisms for sophisticated axon guidance and synaptic plasticity (neurotrophin signalling system; *Figure 7*; Supplementary File 1–Supplementary Table 22, Supplementary File 1–Supplementary Table 23, Supplementary File 1–Supplementary Table 24). These findings are consistent with the convergent evolution of muscle and nerve cells in ctenophores (Moroz *et al.*, 2014) and suggest that bilaterians have a common genetic basis for nervous system patterning despite the recently proposed scenario of convergent evolution of bilaterian nerve cords (Martín-Durán *et al.*, 2018). The importance of the nervous system-related category of bilaterian-specific genes is further underscored by the identification of various transcription factors with a well supported role in nervous system development across phyla, e. g. the Prospero homeobox protein, the Achaete-scute homolog 2, or the neuronal PAS domain-containing protein 4 (Supplementary File 1–Supplementary Table 10, Supplementary File 4). Further, three transmembrane proteins with expression in the nervous system, but unknown function, provide the opportunity to characterize novel factors with nervous system-related function (Supplementary File 1–Supplementary Table 11). Together, the factors we found in this category provide fundamental features of bilaterian nervous systems, and their evolutionary origin in the bilaterian ancestor is compatible with observable changes in nervous system development and architecture.

An unexpectedly high number of bilaterian-specific genes has transcription factor activity (*Figure 3B*; *Figure 2*). As noted above, these factors are often equipped with multiple C₂H₂ zinc fin-

ger domains (**Figure 2–Figure Supplement 1**, Supplementary File 1–Supplementary Table 8). Apart from so far uncharacterized proteins, which include ZF64B_HUMAN or ZN236_HUMAN, the expression and developmental role of bilaterian-specific zinc finger proteins is compatible with prominent functions during early development, such as imaginal disc development (Rotund; *St Pierre et al. (2002)*), modulation of TGF- β signalling (Schnurri; *Yao et al. (2006)*), nephron patterning (Evi1, odd-skipped related 1; *Mead et al. (2005)*; *Dressler (2006)*; *Tena et al. (2007)*; *Li et al. (2014)*), or the differentiation of cardiac precursor cells at the ventral midline (Castor; *Christine and Conlon (2008)*). Importantly, the identified transcription factors with homeobox or bHLH domain are involved in the specification of several bilaterian tissues, the mesoderm (MyoD, PRRX1_HUMAN, BHE22_HUMAN), the nervous system (Prospero homeobox protein 2, Achaete-scute homolog 2, FER3L_HUMAN, NPAS4_HUMAN, BHE22_HUMAN, BUN1_DROME), or the intestine (ISX_HUMAN) (Supplementary File 1–Supplementary Table 10), consistent with a role in the evolution of these characteristic bilaterian traits.

A contiguous cluster of at least seven Hox genes is an ancestral bilaterian feature (*Baguña et al., 2008*). A prerequisite for its formation is the existence of anterior, central, and posterior Hox genes. Our results confirm previous findings that placed the origin of central Hox genes to the bilaterian ancestor (Supplementary File 1–Supplementary Table 10), in contrast to evolutionarily older anterior and posterior Hox genes (*Finnerty and Martindale, 1999*; *Hueber et al., 2013*). Importantly, Hox gene expression is regulated in part by the chromatin organizer CTCF (*Rousseau et al., 2014*; *Narendra et al., 2015*), another bilaterian-specific protein (*Heger et al. (2012)*; Supplementary File 1–Supplementary Table 8; Supplementary File 4). As outlined elsewhere, the evolution of CTCF—and other poly-zinc finger proteins—could have provided a mechanism for the creation and regulation of bilaterian Hox gene clusters, once central Hox genes had been added to the repertoire (*Heger et al., 2012*).

The emergence of the mesoderm as a third germ layer is one of the most characteristic morphological innovations of bilaterian animals. In contrast to previous work, our findings suggest that several genes and gene networks which provide regulatory inputs to mesodermal patterning arose in the bilaterian ancestor. Specifically, we identified orthologs of all Nodal pathway members across bilaterians, but not outside this clade (**Figure 5–Figure Supplement 1**, **Figure 5–Figure Supplement 2**, **Figure 5–Figure Supplement 3**, **Figure 5–Figure Supplement 4**; Supplementary File 1–Supplementary Table 14; Supplementary File 1–Supplementary Table 16). The robust bilaterian-specific distribution of these genes, derived from orthology clustering and HMM-HMM searches, implies that the entire Nodal pathway—and its roles in mesoderm specification and left-right asymmetry—is a bilaterian novelty (**Figure 5**). Although a reasonable speculation, this is currently not supported for all pathway members by phylogenetic analyses and needs to be tested more thoroughly in the future. Together with the bilaterian specificity of additional modulators and effectors of Nodal and/or TGF- β signalling (BAMBI_HUMAN, VWC2_HUMAN, MECOM_HUMAN, Q24605_DROME; Supplementary File 4), these findings suggest that significant changes in TGF- β signalling occurred in the bilaterian ancestor. In addition to the Nodal pathway, several other genes with key roles in mesoderm formation also originated in the bilaterian ancestor, among them the master regulator of muscle cell specification, MyoD, and the Paired mesoderm homeobox protein 1 (PRRX1_HUMAN; Supplementary File 1–Supplementary Table 10) which regulates the formation of preskeletal condensations from undifferentiated mesenchyme during mouse skeletogenesis (*Martin et al., 1995*). Taken together, we identified multiple genetic factors essential for the differentiation of mesoderm and mesodermal tissues in bilaterians.

In conclusion, we demonstrate that a considerable number of genes has a bilaterian-specific distribution and probably originated in the bilaterian ancestor. While the function of some of these genes is unknown, many of them participate in the formation of key morphological innovations in extant bilaterians, implying that the evolution of specific genes contributed to the formation of bilaterian body plans.

Methods and Materials

Sequence collection and database construction

The sequence repertory for this study was assembled from three parts. Genomic and transcriptomic sequences were collected from the sources listed in Supplementary File 1–Supplementary Table 1, Supplementary File 1–Supplementary Table 3, Supplementary File 2. As third component, selected sequences were downloaded from the NCBI non-redundant protein database.

The 25 genomic sequences were first screened for repetitive sequence content using RepeatMasker V4.0.5 (<http://repeatmasker.org>) with default parameters. The resulting contigs/scaffolds were translated into six open reading frames (ORFs) using the Emboss tool «getorf» (Rice et al., 2000), with a minimum ORF length of 25 AA. Sequences containing strings of «X» characters, a result of translating sequencing gaps and masked repeats, were treated differentially to retain as much information as possible. Sequences with ≥ 9 «X» in a row were split. After removing the Xs, each flanking region ≥ 35 valid amino acids was kept and given a new identifier while smaller flanking regions were discarded. These measures decreased sequence count by 46.8 %, from 324,788,561 to 172,606,165 ORFs. To further reduce the amount of ORFs, we blasted them against a custom database of opisthokont sequences. This database contained all sequences of opisthokont origin as extracted from the non-redundant protein database at GenBank, release 198 from October 21, 2013 (2,695,641 sequences). We kept ORFs with a Blast expectation value < 10 against this database and thus rejected ORFs that have no detectable similarity to the protein repertoire of opisthokonts. In a final step, we used CD-HIT (Li and Godzik, 2006) with default parameters and 90 % identity threshold to remove redundancy. These steps reduced the number of sequences significantly, from initially 324,788,561 to 109,567,344 genomic ORFs.

To fill in the gaps of public sequence repositories and extend coverage, we collected transcriptome data of poorly represented animal groups (Supplementary File 1–Supplementary Table 1, Supplementary File 2). Downloaded transcriptomes were first assayed for completeness using the CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline which reports the coverage of 248 ultra-conserved core eukaryotic genes present in a dataset (Parra et al., 2007). On the basis of CEGMA completeness and phylogenetic placement, we selected transcriptomes of 64 species for the dataset. Their average transcriptome completeness according to CEGMA was 60.8 %, with several bilaterian and non-bilaterian species exceeding 90 % (Supplementary File 1–Supplementary Table 2). As described for genomes, transcriptomes were then translated into six open reading frames. We kept the three longest ORFs for each transcriptome contig, removed Xs, and obtained 11,768,516 transcriptome protein sequences in total (Table 2).

To provide a backbone of published and annotated protein sequences for the genomic and transcriptomic ORFs, we filtered the NCBI non-redundant protein database and kept 2.9 million protein sequences from 204 opisthokont species that had $>8,000$ sequence entries each. Extraction of opisthokont sequences was guided by NCBI taxonomy.

As the combination of sequences from three sources again introduced redundancy, we clustered the final dataset with 90 % identity threshold. In a last pre-processing step, we changed the headers of all sequences to obey a consistent naming scheme. It includes the NCBI taxon identifier and a unique sequence ID that allows to distinguish between NCBI-, ORF-, and transcriptome-derived sequences. The final dataset used for this analysis contained 124,031,501 sequences.

Orthology pipeline and clustering

For orthology clustering, we employed the OrthoMCL pipeline (Li et al., 2003). It utilises a graph-based clustering approach for the generation of orthologous groups on the basis of normalised BLAST similarity measurements between sequence pairs. To enable the processing of our large dataset, we ported to the statistical programming environment R (<https://www.r-project.org/>) all steps of the original OrthoMCL pipeline that require interaction with a MySQL database. In this way, loading of the database and inference of orthology tables is limited only by the size of the com-

puter's main memory, not by the speed and additional memory requirements of the underlying MySQL engine, as in the original implementation. By dividing the computation of orthology tables into an appropriate number of steps, our entire dataset could be processed on a compute server with 250 GB memory. Importantly, the R version of OrthoMCL accurately reproduces all steps of the original pipeline (Supplementary File 1–Supplementary Table 4). The collection of scripts for the R version of OrthoMCL is available at <https://github.com/prheger/BigWenDB>.

HMM-HMM searches and database

We extracted from the BigWenDB sequence collection the individual sequences belonging to each of the 824,605 ortholog groups and calculated 824,605 corresponding multiple sequence alignments using default parameters of the MAFFT v7.304b «einsi» algorithm (Kato et al., 2005). After converting the alignments into hhm format (hhsearch format for hidden Markov models) with the command «hhmake» and default parameters, we concatenated them to a database that can be searched by hhsearch (parameters in addition to default: «-nodssp -nopred -dbstrlen 100»), according to Soding (2005). We precomputed HMM-HMM search results for about 20% of orthogroups and issued missing searches on demand. Reciprocal best hit relationships were analysed using custom scripts.

Quality control of clusterings and the bilaterian-specific gene set

Quality control of clustering results and the bilaterian-specific gene set was carried out as described in Appendix 1, sections «Cluster evaluation and quality control» and «Identification of bilaterian-specific genes».

Statistical tests for the enrichment of transcription factors

To test whether the bilaterian-specific gene set of 157 orthogroups is enriched for transcription factors, we downloaded as control the human proteome with 20,205 protein sequences from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/ and predicted transcription factors in this dataset using the PfamScan software (<ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/>) with E value cutoff = 0.00005. We then determined the abundance of ten prevalent DNA-binding domains in the dataset: «Basic; bZIP_2; HLH; HNF-1_N; Homeobox; Hox9_act; HPD; SOBP; THAP; zf-». Corresponding domains were identified in 1,756 of the 20,205 human reference proteins. We then randomly selected 10× 157 genes from the reference set and specified the number of transcription factors (proteins with the above mentioned domains) in the obtained subsets. While the average number of transcription factors in the ten control sets was 12.8 ± 4.44 , the equally sized bilaterian-specific gene set (157 orthogroups) had 37 transcription factors. Modelling a normal distribution from the obtained mean and standard deviation yielded a p -value of $2.512e^{-08}$ for the transcription factor content in bilaterian-specific genes (using the R function «pnorm»). Likewise, a Pearson's χ^2 test with the corresponding data matrix (1,765:20205; 37:157), using the R function «chisq.test», yielded a p -value of $3.805e^{-08}$. Finally, under the assumption of a binomial distribution (R function «pbinom») and given that there are 1,756 transcription factors in 20,205 human proteins, the probability that we obtain 36 or more transcription factors when drawing 157 random proteins is $p < 1.841e^{-08}$.

Poly-Zinc finger scan across Opisthokonta

We downloaded the proteomes of seven ecdysozoan, five lophotrochozoan, 12 deuterostomian, and four non-bilaterian species from <http://www.uniprot.org/proteomes>. On average, each proteome consisted of 28,772 sequences. We scanned all protein sequences for the presence of protein domains using the PfamScan software (<ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/>) with E value cutoff = 0.00005 and Pfam database version 31.0. Using command line tools, we identified C₂H₂ zinc finger proteins in the PfamScan output and counted for every proteome the number of pro-

teins with six or more zinc finger domains. The resulting numbers were used to plot **Figure 2–Figure Supplement 1A, B**.

To determine the number of poly-ZF proteins that originated in the ancestor of opisthokonts, metazoans, and eumetazoans, we first extracted from the clustering results orthogroups specific for these lineages. The filtering criteria for selecting opisthokont-specific orthogroups were: Fungi ≥ 20 species, Metazoa ≥ 40 species, Bilateria ≥ 30 species and yielded 2,928 orthogroups of ancient origin. The filtering criteria for selecting metazoan-specific orthogroups were identical, except that no fungi were allowed, and yielded 2,615 metazoan-specific orthogroups. For eumetazoan-specific orthogroups we required the presence of at least 30 bilaterian and 3 cnidarian species, with not more than 2 ctenophore species allowed (according to NCBI taxonomy, both ctenophores and cnidarians misleadingly belong to eumetazoans). Applying these conditions, we obtained 283 eumetazoan-specific orthogroups. Next we extracted the longest sequence of each opisthokont-, metazoan-, and eumetazoan-specific orthogroup and scanned it with PfamScan (E value cutoff = 0.00005). Finally, we counted the number of poly-ZF sequences with at least six domains for each node and mapped the numbers to a phylogeny. Note that this «simple» filtering strategy (Bilateria: ≥ 30 species) would return 662 bilaterian-specific orthogroups, considerably more than the 157 error-corrected orthogroups in the final dataset. The strategy therefore possibly overestimates the number of poly-ZF proteins at the three ancient nodes.

Determining orthogroup ancestors

To determine the ancestor of the species combined in a given orthogroup, we wrote a custom Perl script that extracts the taxonomic identifiers of each sequence and then determines the last common ancestor of all represented species on the basis of NCBI taxonomy and lineage information (<ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/>). The script generates output that can be parsed and filtered using command line utilities. It is part of the collection of R scripts at <https://github.com/prheger/BigWenDB>.

Protein domain scans and gene ontology analysis

We applied strict filtering rules to extract bilaterian-, vertebrate-, and arthropod-specific genes from the Markov clustering results (rule for bilaterian-specific orthogroups: deuterostomes ≥ 7 , lophotrochozoans ≥ 4 or 0, ecdysozoans ≥ 4 or 0; for arthropod-specific orthogroups: chelicerates ≥ 2 , crustaceans ≥ 0 , myriapods ≥ 1 , insects ≥ 5 ; for vertebrate-specific orthogroups: ≥ 40 of 53 gnathostome species). From each lineage-specific orthogroup obtained by filtering, we extracted the longest sequence and scanned it with PfamScan Version 1.5 (*Punta et al., 2012*) (available at <ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/>) at an *E-value* cutoff of e^{-05} for the presence of protein domains as classified in PFAM database release 30.0 (release date: 06/16).

To associate the identified protein domains with gene ontology (GO) terms, we utilised the Pfam2GO list at <http://geneontology.org/external2go/pfam2go> and extracted relevant terms using command line tools. Typically, only a subset of domains was linked to GO terms. We finally created a list with the relative number of identified protein domains and associated gene ontology terms and visualized this list as a word cloud at www.wortwolken.com.

Multiple sequence alignment and phylogenetic analysis

Multiple sequence alignments required for the HMM-HMM database and phylogenetic analyses were carried out using the MAFFT v7.304b «einsi» algorithm with default parameters (*Katoh et al., 2005*). Large alignments (> 200 sequences) were computed using MAFFT v7.304b with high speed parameters. For phylogeny, we added ingroup and outgroup sequences from the clustered orthogroup sets or from public repositories, as appropriate, and manually removed indels and unalignable regions from the data prior to analysis. In some cases, e. g. for Lefty, we generated a hidden Markov model of an orthogroup alignment and searched additional transcriptomic datasets

not represented in the BigWenDB for potential orthologs. Phylogenetic trees were computed under the maximum likelihood criterion, using IQ-TREE v1.6.10 (Nguyen *et al.*, 2015) with ModelFinder for fast and accurate model selection (Kalyaanamoorthy *et al.*, 2017), ultrafast bootstrap approximation and optimization (1,000 replicates) (Minh *et al.*, 2013), and Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-aLRT) (command line parameters: -bb 1000 -alrt 1000 -bnni). Resulting trees were edited with FigTree v1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>) and Affinity Designer Version 1.72 (<https://affinity.serif.com>).

Prediction of protein structure

After constructing multiple sequence alignments from cnidarian and bilaterian Robo proteins, we identified the transmembrane region (corresponding to sequence «AFIAGIGAACWIIIMVFSIWL» in ROBO1_HUMAN) and generated two subsequences overlapping at this feature. One subsequence spanned the extracellular part of the protein plus the transmembrane domain, the other spanned the transmembrane domain plus the cytoplasmic part. We generated the two fragments for seven exemplary Robo proteins, for the deuterostomes *Homo sapiens* and *Strongylocentrotus purpuratus*, the lophotrochozoan *Lingula anatina*, the ecdysozoans *Drosophila melanogaster* and *Trichinella pseudospiralis*, and the two cnidarians *Hydra vulgaris* and *Stylophora pistillata*. All fragments were uploaded to the Phyre2 web interface (<http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index>; (Kelley *et al.*, 2015)) and analysed with modelling mode «intensive» (complete modelling using multiple templates and ab initio techniques).

Identification of metazoan-specific genes

To obtain a list of genes with metazoan origin, we first blasted 20,205 human genes obtained from uniprot.org against the BigWen database and obtained Blast hits for 19,322 genes. To reliably map the UniProt queries to orthogroups, we selected queries that had a Blast hit with high identity (>95 percent) over at least 100 amino acids. For proteins fulfilling these criteria, we extracted the corresponding orthogroup ID and ancestor, taking into account only orthogroups with at least 75 species to ensure broad sampling. After removing redundancy, we obtained 797 distinct orthogroups of metazoan origin whose human orthologs were used for the stringDB PPI network analysis. A conceptually similar study obtained 1,189 novel metazoan-specific homology groups, which is in reasonable agreement with our result when considering the differences in methodology and datasets (Paps and Holland, 2018).

Protein-protein interaction network analyses

Protein interaction data were obtained from the STRING database v.11.0 of known and predicted protein-protein interactions (PPI; <https://string-db.org>; Szklarczyk *et al.* (2017)). To construct PPI networks, we first identified the appropriate human orthologs of bilaterian-specific and metazoan-specific orthogroups. We obtained 150 human ortholog IDs for the 157 bilaterian-specific orthogroups and 797 human ortholog IDs for 797 metazoan-specific orthogroups (collected as described above). We uploaded these protein IDs to the STRING browser interface and generated three separate PPI networks, one for bilaterian-specific proteins (B), one for metazoan-specific proteins (M), and a combined network for both taxonomic groups (B + M). The average local clustering coefficients and PPI enrichment p-values we report are based on analyses with default settings, where all evidence types were considered. Further statistical analyses were conducted for the B + M full network and the B + M Nodal-Lefty subnetwork, the latter being defined by the core five bilaterian-specific proteins (Nodal, Lefty, FoxH1, Eomes, and EGF-CFC) and their interaction partners. From the complete list of pairwise protein-protein interactions in the B + M network, data were extracted for the numbers of B – B, M – M, and B – M interactions and assessed by a χ^2 test. Additional calculations were made per protein for the total number of interactions and for the proportion of these that involve a bilaterian-specific interaction partner. Boxplots for these values display the median, and whiskers represent 1.5x the value of the Q3 (upper) or Q2 (lower) quartile range, with

outliers omitted for clarity. Statistical tests involved χ^2 tests (<https://www.socscistatistics.com/tests/chisquare/default2.aspx>, accessed 26 August 2019) and non-parametric comparisons in multigroup (Kruskal-Wallis) and pairwise (Mann-Whitney U) assessments as reported, calculated in R version 3.4.0 and from the Python library scipy.stats (function: mannwhitneyu).

Data Access

The R version of OrthoMCL and a script for inferring orthogroup ancestors are available at <https://github.com/prheger/BigWenDB>. The sequence dataset used to build the BigWenDB and the final clustering results are available at <https://datadryad.org/review?doi=doi:10.5061/dryad.4qf7168>. Several Supplementary Files with original data and Supplementary Tables are linked to this paper at elifesciences.org:

1. File S1 — Supplementary Tables 1–26 and Supplementary Figures 1 and 2: supplementary_file_tables.pdf (.pdf document)
2. File S2 — Download location for transcriptome data used in this study: data_availability_previously_published_datasets_v2.xls (spreadsheet in .xls format)
3. File S3 — Comparison between Orthobench and BigWenDB clustering results: orthobench_comparison_result_wen.tsv (tab-delimited file)
4. File S4 — List of high-confidence bilaterian-specific orthogroups: 157_bilat-spec_OGs_infos.sorted_ncbi-blast-anno4_+DmV2.ods (spreadsheet in .ods format)

Acknowledgments

This research was supported by grants from the German Research Foundation to TW (CRC 680 and CRC 1211) and to KAP (CRC 680). BLAST searches were computed on CHEOPS, the Cologne High Efficiency Operating Platform for Science of the University of Cologne, and on JuRoPA (Jülich Research on Petaflop Architectures), a High Performance Computing Platform of the Jülich Supercomputing Centre, Germany. We thank Robert Fürst for programming help, Kay Hofmann for help with protein structure analysis, Richard Stancliffe for scripting and statistical support, Maria Thieser for help with transcriptome processing, and Olav Zimmermann for the cooperation with the Jülich Supercomputing Centre. Special thanks to countless researchers and institutions for sharing sequence data.

References

- Acemel RD**, Maeso I, Gómez-Skarmeta JL. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdiscip Rev Dev Biol*. 2017 May; 6. doi: 10.1002/wdev.265.
- Afroze S**, Meng F, Jensen K, McDaniel K, Rahal K, Onori P, Gaudio E, Alpini G, Glaser SS. The physiological roles of secretin and its receptor. *Annals of Translational Medicine*. 2013 Oct; 1:29. doi: 10.3978/j.issn.2305-5839.2012.12.01.
- Aguilera F**, McDougall C, Degnan BM. Co-Option and De Novo Gene Evolution Underlie Molluscan Shell Diversity. *Molecular Biology and Evolution*. 2017 Apr; 34:779–792. doi: 10.1093/molbev/msw294.
- Aguileta G**, Marthey S, Chiapello H, Lebrun MH, Rodolphe F, Fournier E, Gendault-Jacquemard A, Giraud T. Assessing the performance of single-copy genes for recovering robust phylogenies. *Systematic Biology*. 2008 Aug; 57:613–627. doi: 10.1080/10635150802306527.
- Aguinaldo AM**, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*. 1997; 387(6632):489–93. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=9168109>.
- Anttil M**. Chemical transmission in the sea anemone *Nematostella vectensis*: A genomic perspective. *Comp Biochem Physiol Part D Genomics Proteomics*. 2009 Dec; 4(4):268–289. <http://dx.doi.org/10.1016/j.cbd.2009.07.001>, doi: 10.1016/j.cbd.2009.07.001.

- 990 **Arenas-Mena C.** The transcription factors HeBlimp and HeT-brain of an indirectly developing polychaete suggest ancestral endodermal, gastrulation, and sensory cell-type specification roles. *Journal of Experimental Zoology Part B, Molecular and Developmental Evolution*. 2008 Nov; 310:567–576. doi: [10.1002/jez.b.21225](https://doi.org/10.1002/jez.b.21225).
- 993 **Arnold SJ, Hofmann UK, Bikoff EK, Robertson EJ.** Pivotal roles for eomesodermin during axis formation, epithelium-to-mesenchyme transition and endoderm specification in the mouse. *Development*. 2008 Feb; 135:501–511. doi: [10.1242/dev.014357](https://doi.org/10.1242/dev.014357).
- 996 **Babonis LS, Martindale MQ.** Phylogenetic evidence for the modular evolution of metazoan signalling pathways. *Philosophical Transactions of the Royal Society of London*. 2017 Feb; 372. doi: [10.1098/rstb.2015.0477](https://doi.org/10.1098/rstb.2015.0477).
- 998 **Baguña J, Martínez P, Paps J, Riutort M.** Back in time: a new systematic proposal for the Bilateria. *Philos Trans R Soc Lond B Biol Sci*. 2008; 363(1496):1481–91. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18192186>.
- 1001 **Balavoine G, de Rosa R, Adoutte A.** Hox clusters and bilaterian phylogeny. *Mol Phylogenet Evol*. 2002; 24(3):366–73. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=12220978>.
- 1004 **Barron AB, Søvik E, Cornish JL.** The roles of dopamine and related compounds in reward-seeking behavior across animal phyla. *Frontiers in Behavioral Neuroscience*. 2010; 4:163. doi: [10.3389/fnbeh.2010.00163](https://doi.org/10.3389/fnbeh.2010.00163).
- 1006 **Bashaw GJ, Kidd T, Murray D, Pawson T, Goodman CS.** Repulsive axon guidance: Abelson and Enabled play opposing roles downstream of the roundabout receptor. *Cell*. 2000 Jun; 101:703–715.
- 1008 **Bauknecht P, Jékely G.** Ancient coexistence of norepinephrine, tyramine, and octopamine signaling in bilaterians. *BMC Biology*. 2017 Jan; 15:6. doi: [10.1186/s12915-016-0341-7](https://doi.org/10.1186/s12915-016-0341-7).
- 1010 **Benito-Gutiérrez E, Nake C, Llovera M, Comella JX, Garcia-Fernández J.** The single AmphiTrk receptor highlights increased complexity of neurotrophin signalling in vertebrates and suggests an early role in developing sensory neuroepidermal cells. *Development*. 2005 May; 132:2191–2202. doi: [10.1242/dev.01803](https://doi.org/10.1242/dev.01803).
- 1013 **Berger M, Gray JA, Roth BL.** The expanded biology of serotonin. *Annual Review of Medicine*. 2009; 60:355–366. doi: [10.1146/annurev.med.60.042307.110802](https://doi.org/10.1146/annurev.med.60.042307.110802).
- 1015 **Berridge KC.** Motivation concepts in behavioral neuroscience. *Physiology & Behavior*. 2004 Apr; 81:179–209. doi: [10.1016/j.physbeh.2004.02.004](https://doi.org/10.1016/j.physbeh.2004.02.004).
- 1017 **Bosch TCG, Klimovich A, Domazet-Lošo T, Gründer S, Holstein TW, Jékely G, Miller DJ, Murillo-Rincon AP, Rentzsch F, Richards GS, Schröder K, Technau U, Yuste R.** Back to the Basics: Cnidarians Start to Fire. *Trends in Neurosciences*. 2017 Feb; 40:92–105. doi: [10.1016/j.tins.2016.11.005](https://doi.org/10.1016/j.tins.2016.11.005).
- 1020 **Brooke NM, Holland PWH.** The evolution of multicellularity and early animal genomes. *Curr Opin Genet Dev*. 2003 Dec; 13(6):599–603.
- 1022 **Brose K, Bland KS, Wang KH, Arnott D, Henzel W, Goodman CS, Tessier-Lavigne M, Kidd T.** Slit proteins bind Robo receptors and have an evolutionarily conserved role in repulsive axon guidance. *Cell*. 1999 Mar; 96:795–806.
- 1025 **Brown KM, Shaver JR.** [3H]Serotonin binding to blastula, gastrula, prism, and pluteus sea urchin embryo cells. *Comparative Biochemistry and Physiology*. 1989; 93C:281–285. doi: [10.1016/0742-8413\(89\)90234-x](https://doi.org/10.1016/0742-8413(89)90234-x).
- 1027 **Budd GE.** The earliest fossil record of the animals and its significance. *Philos Trans R Soc Lond B Biol Sci*. 2008; 363(1496):1425–34. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18192192>.
- 1030 **Budd GE, Jensen S.** The origin of the animals and a ‘Savannah’ hypothesis for early bilaterian evolution. *Biological Reviews of the Cambridge Philosophical Society*. 2017 Feb; 92:446–473. doi: [10.1111/brv.12239](https://doi.org/10.1111/brv.12239).
- 1032 **Budhiraja S, Chugh A.** Neuromedin U: physiology, pharmacology and therapeutic potential. *Fundamental & Clinical Pharmacology*. 2009 Apr; 23:149–157. doi: [10.1111/j.1472-8206.2009.00667.x](https://doi.org/10.1111/j.1472-8206.2009.00667.x).
- 1034 **Burke CJ, Huetteroth W, Oswald D, Perisse E, Krashes MJ, Das G, Gohl D, Silies M, Certel S, Waddell S.** Layered reward signalling through octopamine and dopamine in *Drosophila*. *Nature*. 2012 Dec; 492:433–437. doi: [10.1038/nature11614](https://doi.org/10.1038/nature11614).

- 1037 **Buznikov GA**, Lambert HW, Lauder JM. Serotonin and serotonin-like substances as regulators of
1038 early embryogenesis and morphogenesis. *Cell and Tissue Research*. 2001 Aug; 305:177–186. doi:
1039 10.1007/s004410100408.
- 1040 **Canfield DE**, Poulton SW, Narbonne GM. Late-Neoproterozoic deep-ocean oxygenation and the rise of ani-
1041 mal life. *Science*. 2007; 315(5808):92–5. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17158290)
1042 [dbfrom=pubmed&retmode=ref&id=17158290](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17158290).
- 1043 **Cannon JT**, Vellutini BC, Smith J 3rd, Ronquist F, Jondelius U, Hejnol A. Xenacoelomorpha is the sister group
1044 to Nephrozoa. *Nature*. 2016 Feb; 530(7588):89–93. <http://dx.doi.org/10.1038/nature16520>, doi: 10.1038/na-
1045 ture16520.
- 1046 **Cardoso JCR**, Félix RC, Bergqvist CA, Larhammar D. New insights into the evolution of vertebrate CRH
1047 (corticotropin-releasing hormone) and invertebrate DH44 (diuretic hormone 44) receptors in metazoans.
1048 *General and Comparative Endocrinology*. 2014 Dec; 209:162–170. doi: 10.1016/j.ygcen.2014.09.004.
- 1049 **Cardoso JCR**, Pinto VC, Vieira FA, Clark MS, Power DM. Evolution of secretin family GPCR members in the
1050 Metazoa. *BMC Evol Biol*. 2006 Dec; 6:108. doi: 10.1186/1471-2148-6-108.
- 1051 **Carlberg M**, Anctil M. Biogenic amines in coelenterates. *Comparative Biochemistry and Physiology C, Compar-*
1052 *ative Pharmacology and Toxicology*. 1993 Sep; 106:1–9.
- 1053 **Carroll SB**. Homeotic genes and the evolution of arthropods and chordates. *Nature*. 1995; 376(6540):479–
1054 85. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=7637779)
1055 [7637779](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=7637779).
- 1056 **Chan C**, Jayasekera S, Kao B, Páramo M, von Grotthuss M, Ranz JM. Remodelling of a homeobox gene
1057 cluster by multiple independent gene reunions in *Drosophila*. *Nat Commun*. 2015 Mar; 6:6509. doi:
1058 10.1038/ncomms7509.
- 1059 **Chen F**, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to
1060 eukaryotic genomes. *PLoS ONE*. 2007 Apr; 2:e383. doi: 10.1371/journal.pone.0000383, original DateCom-
1061 pleted: 20070730.
- 1062 **Chen L**, Krause M, Sepanski M, Fire A. The *Caenorhabditis elegans* MYOD homologue HLH-1 is essential for
1063 proper muscle function and complete morphogenesis. *Development*. 1994 Jun; 120:1631–1641.
- 1064 **Chen Y**, Schier AF. Lefty proteins are long-range inhibitors of squint-mediated nodal signaling. *Curr Biol*. 2002
1065 Dec; 12:2124–2128.
- 1066 **Chourrout D**, Delsuc F, Chourrout P, Edvardsen RB, Rentzsch F, Renfer E, Jensen MF, Zhu B, de Jong P, Steele
1067 RE, Technau U. Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature*.
1068 2006; 442(7103):684–7. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16900199)
1069 [retmode=ref&id=16900199](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16900199).
- 1070 **Christine KS**, Conlon FL. Vertebrate CASTOR is required for differentiation of cardiac precursor cells at the
1071 ventral midline. *Dev Cell*. 2008 Apr; 14:616–623. doi: 10.1016/j.devcel.2008.01.009.
- 1072 **Conant GC**, Wagner A. Asymmetric sequence divergence of duplicate genes. *Genome Research*. 2003 Sep;
1073 13:2052–2058. doi: 10.1101/gr.1252603.
- 1074 **Conway Morris S**. Darwin's dilemma: the realities of the Cambrian 'explosion'. *Philos Trans R Soc Lond B Biol*
1075 *Sci*. 2006; 361(1470):1069–83. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16754615)
1076 [pubmed&retmode=ref&id=16754615](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16754615).
- 1077 **Darwin C**. *The Origin of Species: By Means of Natural Selection, or the Preservation of Favoured Races in*
1078 *the Struggle for Life*. 6th ed. Cambridge Library Collection - Darwin, Evolution and Genetics, Cambridge
1079 University Press; 2009. doi: 10.1017/CBO9780511694295.
- 1080 **Daubert EA**, Condron BG. Serotonin: a regulator of neuronal morphology and circuitry. *Trends in Neuro-*
1081 *sciences*. 2010 Sep; 33:424–434. doi: 10.1016/j.tins.2010.05.005.
- 1082 **Davidson EH**, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science*. 2006;
1083 311(5762):796–800. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16469913)
1084 [retmode=ref&id=16469913](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16469913).

- 1085 **Desgrange A**, Cereghini S. Nephron Patterning: Lessons from Xenopus, Zebrafish, and Mouse Studies. *Cells*.
1086 2015 Sep; 4:483–499. doi: 10.3390/cells4030483.
- 1087 **van Dongen S**. Graph Clustering by Flow Simulation. PhD thesis,. 2000; University of Utrecht.
- 1088 **Dressler GR**. The cellular basis of kidney development. *Annual Review of Cell and Developmental Biology*.
1089 2006; 22:509–529. doi: 10.1146/annurev.cellbio.22.010305.104340.
- 1090 **Duboc V**, Lapraz F, Besnardeau L, Lepage T. Lefty acts as an essential modulator of Nodal activity during sea
1091 urchin oral-aboral axis formation. *Dev Biol*. 2008 Aug; 320:49–59. doi: 10.1016/j.ydbio.2008.04.012.
- 1092 **Duboc V**, Röttinger E, Besnardeau L, Lepage T. Nodal and BMP2/4 signaling organizes the oral-aboral axis of
1093 the sea urchin embryo. *Dev Cell*. 2004 Mar; 6:397–410.
- 1094 **Dunn CW**, Hejnal A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD,
1095 Sorensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet
1096 G. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 2008; 452(7188):745–
1097 9. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18322464)
1098 [18322464](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18322464).
- 1099 **Dunn CW**, Giribet G, Edgecombe GD, Hejnal A. Animal phylogeny and its evolutionary implications. *Annual*
1100 *Review of Ecology, Evolution, and Systematics*. 2014 Submitted, 9 Feb; 45:371–395.
- 1101 **Dunwell TL**, Paps J, Holland PWH. Novel and divergent genes in the evolution of placental mammals. *Proceed-*
1102 *ings Biological Sciences*. 2017 Oct; 284. doi: 10.1098/rspb.2017.1357.
- 1103 **Duttke SHC**, Doolittle RF, Wang YL, Kadonaga JT. TRF2 and the evolution of the Bilateria. *Genes Dev*. 2014 Oct;
1104 28(19):2071–2076. <http://dx.doi.org/10.1101/gad.250563.114>, doi: 10.1101/gad.250563.114.
- 1105 **Eddy SR**. Profile hidden Markov models. *Bioinformatics*. 1998; 14(9):755–63. [http://eutils.ncbi.nlm.nih.gov/](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=9918945)
1106 [entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=9918945](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=9918945).
- 1107 **El-Shehabi F**, Taman A, Moali LS, El-Sakkary N, Ribeiro P. A novel G protein-coupled receptor of *Schistosoma*
1108 *mansoni* (SmGPR-3) is activated by dopamine and is widely expressed in the nervous system. *PLoS Neglected*
1109 *Tropical Diseases*. 2012; 6:e1523. doi: 10.1371/journal.pntd.0001523.
- 1110 **Emerson RO**, Thomas JH. Adaptive evolution in zinc finger transcription factors. *PLoS Genet*. 2009;
1111 5(1):e1000325. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19119423)
1112 [ref&id=19119423](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19119423).
- 1113 **Emms DM**, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramati-
1114 cally improves orthogroup inference accuracy. *Genome Biol*. 2015; 16:157. [http://dx.doi.org/10.1186/](http://dx.doi.org/10.1186/s13059-015-0721-2)
1115 [s13059-015-0721-2](http://dx.doi.org/10.1186/s13059-015-0721-2), doi: 10.1186/s13059-015-0721-2.
- 1116 **Erwin DH**, Valentine JW. The Cambrian explosion: The construction of animal biodiversity. Roberts and Com-
1117 pany Publishers, Inc., Greenwood Village, USA; 2013.
- 1118 **Evans TA**. Embryonic axon guidance: insights from *Drosophila* and other insects. *Current Opinion in Insect*
1119 *Science*. 2016 Dec; 18:11–16. doi: 10.1016/j.cois.2016.08.007.
- 1120 **Fan X**, Dougan ST. The evolutionary origin of nodal-related genes in teleosts. *Dev Genes Evol*. 2007 Dec;
1121 217:807–813. doi: 10.1007/s00427-007-0191-y.
- 1122 **Feuda R**, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, Wörheide G, Pisani D. Improved Mod-
1123 eling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Curr Biol*. 2017 Dec;
1124 27:3864–3870.e4. doi: 10.1016/j.cub.2017.11.008.
- 1125 **Feulner PGD**, De-Kayne R. Genome evolution, structural rearrangements and speciation. *Journal of Evolution-*
1126 *ary Biology*. 2017 Aug; 30:1488–1490. doi: 10.1111/jeb.13101.
- 1127 **Finnerty JR**, Martindale MQ. Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the
1128 Cnidaria. *Evolution & Development*. 1999; 1:16–23.
- 1129 **Fritzenwanker JH**, Gerhart J, Freeman RM, Lowe CJ. The Fox/Forkhead transcription factor family of the hemi-
1130 chordate *Saccoglossus kowalevskii*. *EvoDevo*. 2014; 5:17. doi: 10.1186/2041-9139-5-17.
- 1131 **Galas L**, Bénard M, Lebon A, Komuro Y, Schapman D, Vaudry H, Vaudry D, Komuro H. Postnatal Migration of
1132 Cerebellar Interneurons. *Brain Sciences*. 2017 Jun; 7. doi: 10.3390/brainsci7060062.

- 1133 **Goyama S**, Yamamoto G, Shimabe M, Sato T, Ichikawa M, Ogawa S, Chiba S, Kurokawa M. Evi-1 is a critical
1134 regulator for hematopoietic stem cells and transformed leukemic cells. *Cell Stem Cell*. 2008 Aug; 3:207–220.
1135 doi: [10.1016/j.stem.2008.06.002](https://doi.org/10.1016/j.stem.2008.06.002).
- 1136 **Grande C**, Martín-Durán JM, Kenny NJ, Truchado-García M, Hejnal A. Evolution, divergence and loss of the
1137 Nodal signalling pathway: new data and a synthesis across the Bilateria. *Int J Dev Biol*. 2014; 58(6-8):521–
1138 532. <http://dx.doi.org/10.1387/ijdb.140133cg>, doi: [10.1387/ijdb.140133cg](https://doi.org/10.1387/ijdb.140133cg).
- 1139 **Hallböök F**, Wilson K, Thorndyke M, Olinski RP. Formation and evolution of the chordate neurotrophin and Trk
1140 receptor genes. *Brain, Behavior and Evolution*. 2006; 68:133–144. doi: [10.1159/000094083](https://doi.org/10.1159/000094083).
- 1141 **Harvey RP**. NK-2 homeobox genes and heart development. *Dev Biol*. 1996 Sep; 178:203–216. doi:
1142 [10.1006/dbio.1996.0212](https://doi.org/10.1006/dbio.1996.0212).
- 1143 **He X**, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in
1144 duplicate gene evolution. *Genetics*. 2005 Feb; 169:1157–1164. doi: [10.1534/genetics.104.037051](https://doi.org/10.1534/genetics.104.037051).
- 1145 **Hecker N**, Sharma V, Hiller M. Convergent gene losses illuminate metabolic and physiological changes in
1146 herbivores and carnivores. *Proceedings of the National Academy of Sciences of the United States of America*.
1147 2019 Feb; 116:3036–3041. doi: [10.1073/pnas.1818504116](https://doi.org/10.1073/pnas.1818504116).
- 1148 **Heger P**, George R, Wiehe T. Successive gain of insulator proteins in arthropod evolution. *Evolution*.
1149 2013; 67(10):2945–2956. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=24094345)
1150 [retmode=ref&id=24094345](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=24094345).
- 1151 **Heger P**, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF and the emergence of
1152 metazoan diversity. *Proc Natl Acad Sci U S A*. 2012; 109(43):17507–12. [http://eutils.ncbi.nlm.nih.gov/entrez/](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23045651)
1153 [eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23045651](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23045651).
- 1154 **Heger P**, Wiehe T. New tools in the box: An evolutionary synopsis of chromatin insulators. *Trends Genet*.
1155 2014 May; 30(5):161–71. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=24786278)
1156 [retmode=ref&id=24786278](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=24786278), doi: <https://doi.org/10.1016/j.tig.2014.03.004>.
- 1157 **Hejnal A**, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguña J, Bailly X, Jondelius
1158 U, Wiens M, Muller WE, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn CW. Assessing the root
1159 of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci*. 2009; 276(1677):4261–70. [http://](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19759036)
1160 eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19759036.
- 1161 **Hinck AP**, Mueller TD, Springer TA. Structural Biology and Evolution of the TGF- β Family. *Cold Spring Harbor*
1162 *Perspectives in Biology*. 2016 Dec; 8(12). doi: [10.1101/cshperspect.a022103](https://doi.org/10.1101/cshperspect.a022103).
- 1163 **Huang EJ**, Reichardt LF. Neurotrophins: roles in neuronal development and function. *Annual Review of Neu-*
1164 *roscience*. 2001; 24:677–736. doi: [10.1146/annurev.neuro.24.1.677](https://doi.org/10.1146/annurev.neuro.24.1.677).
- 1165 **Hudson C**, Yasuo H. Patterning across the ascidian neural plate by lateral Nodal signalling sources. *Develop-*
1166 *ment*. 2005 Mar; 132:1199–1210. doi: [10.1242/dev.01688](https://doi.org/10.1242/dev.01688).
- 1167 **Hueber SD**, Rauch J, Djordjevic MA, Gunter H, Weiller GF, Frickey T. Analysis of central Hox protein types across
1168 bilaterian clades: on the diversification of central Hox proteins from an Antennapedia/Hox7-like protein. *Dev*
1169 *Biol*. 2013 Nov; 383:175–185. doi: [10.1016/j.ydbio.2013.09.009](https://doi.org/10.1016/j.ydbio.2013.09.009).
- 1170 **Huerta-Cepas J**, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn
1171 M, Jensen LJ, von Mering C, Bork P. eggNOG 4.5: a hierarchical orthology framework with improved functional
1172 annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*. 2016 Jan; 44(D1):D286–D293.
1173 <http://dx.doi.org/10.1093/nar/gkv1248>, doi: [10.1093/nar/gkv1248](https://doi.org/10.1093/nar/gkv1248).
- 1174 **Huminiecki L**, Goldovsky L, Freilich S, Moustakas A, Ouzounis C, Heldin CH. Emergence, development and
1175 diversification of the TGF- β signalling pathway within the animal kingdom. *BMC Evol Biol*. 2009 Feb; 9:28.
1176 doi: [10.1186/1471-2148-9-28](https://doi.org/10.1186/1471-2148-9-28).
- 1177 **Irvine SQ**, Warinner SA, Hunter JD, Martindale MQ. A survey of homeobox genes in *Chaetopterus variopedatus*
1178 and analysis of polychaete homeodomains. *Mol Phylogenet Evol*. 1997 Jun; 7(3):331–345. [http://dx.doi.org/](http://dx.doi.org/10.1006/mpev.1997.0407)
1179 [10.1006/mpev.1997.0407](http://dx.doi.org/10.1006/mpev.1997.0407), doi: [10.1006/mpev.1997.0407](https://doi.org/10.1006/mpev.1997.0407).
- 1180 **Jackson DJ**, Thiel V, Wörheide G. An evolutionary fast-track to biocalcification. *Geobiology*. 2010 Jun; 8(3):191–
1181 196. <http://dx.doi.org/10.1111/j.1472-4669.2010.00236.x>, doi: [10.1111/j.1472-4669.2010.00236.x](https://doi.org/10.1111/j.1472-4669.2010.00236.x).

- Jagla K**, Bellard M, Frasch M. A cluster of *Drosophila* homeobox genes involved in mesoderm differentiation programs. *Bioessays*. 2001; 23(2):125–33. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=11169585>.
- Jain A**, Perisa D, Fliedner F, von Haeseler A, Ebersberger I. The Evolutionary Traceability of a Protein. *Genome Biology and Evolution*. 2019 Feb; 11:531–545. doi: 10.1093/gbe/evz008.
- Jayaswal V**, Wong TKF, Robinson J, Poladian L, Jermini LS. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic Biology*. 2014 Sep; 63:726–742. doi: 10.1093/sysbio/syu036.
- Jegla TJ**, Zmasek CM, Batalov S, Nayak SK. Evolution of the human ion channel set. *Combinatorial Chemistry & High Throughput Screening*. 2009 Jan; 12:2–23.
- Johnson EC**, Shafer OT, Trigg JS, Park J, Schooley DA, Dow JA, Taghert PH. A novel diuretic hormone receptor in *Drosophila*: evidence for conservation of CGRP signaling. *The Journal of Experimental Biology*. 2005 Apr; 208:1239–1246. doi: 10.1242/jeb.01529.
- Jondelius U**, Ruiz-Trillo I, Baguña J, Riutort M. The Nemertodermatida are basal bilaterians and not members of the Platyhelminthes. *Zoologica Scripta*. 2002; 31(2):201–215. <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1463-6409.2002.00090.x>, doi: 10.1046/j.1463-6409.2002.00090.x.
- Kalyanamoothy S**, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*. 2017 Jun; 14:587–589. doi: 10.1038/nmeth.4285.
- Kass-Simon G**, Pierobon P. Cnidarian chemical neurotransmission, an updated overview. *Comparative Biochemistry and Physiology Part A, Molecular & Integrative Physiology*. 2007 Jan; 146:9–25. doi: 10.1016/j.cbpa.2006.09.008.
- Kassabov SR**, Choi YB, Karl KA, Vishwasrao HD, Bailey CH, Kandel ER. A single *Aplysia* neurotrophin mediates synaptic facilitation via differentially processed isoforms. *Cell Reports*. 2013 Apr; 3:1213–1227. doi: 10.1016/j.celrep.2013.03.008.
- Katoh K**, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005; 33(2):511–8. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15661851>.
- Kelley LA**, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*. 2015 Jun; 10:845–858. doi: 10.1038/nprot.2015.053.
- Kenny NJ**, Namigai EKO, Dearden PK, Hui JHL, Grande C, Shimeld SM. The Lophotrochozoan TGF- β signalling cassette—diversification and conservation in a key signalling pathway. *Int J Dev Biol*. 2014; 58(6-8):533–549. <http://dx.doi.org/10.1387/ijdb.140080nk>, doi: 10.1387/ijdb.140080nk.
- Kent WJ**. BLAT—the BLAST-like alignment tool. *Genome Res*. 2002 Apr; 12:656–664. doi: 10.1101/gr.229202.
- Ketchesin KD**, Stinnett GS, Seasholtz AF. Corticotropin-releasing hormone-binding protein and stress: from invertebrates to humans. *Stress*. 2017 Sep; 20:449–464. doi: 10.1080/10253890.2017.1322575.
- Kidd T**, Bland KS, Goodman CS. Slit is the midline repellent for the robo receptor in *Drosophila*. *Cell*. 1999 Mar; 96:785–794.
- Kim WK**, Marcotte EM. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS Computational Biology*. 2008 Nov; 4:e1000232. doi: 10.1371/journal.pcbi.1000232.
- Kim Y**, Nirenberg M. *Drosophila* NK-homeobox genes. *Proc Natl Acad Sci U S A*. 1989 Oct; 86(20):7716–7720.
- Kim YJ**, Cecchini KR, Kim TH. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc Natl Acad Sci U S A*. 2011 May; 108:7391–7396. doi: 10.1073/pnas.1018279108.
- Klironomos FD**, Berg J, Collins S. How epigenetic mutations can affect genetic evolution: model and mechanism. *BioEssays*. 2013 Jun; 35:571–578. doi: 10.1002/bies.201200169.
- Kobe B**, Kajava AV. The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*. 2001 Dec; 11:725–732.

- 1230 **Kondrashov FA**, Rogozin IB, Wolf YI, Koonin EV. Selection in the evolution of gene duplications. *Genome*
1231 *Biology*. 2002; 3:RESEARCH0008. doi: 10.1186/gb-2002-3-2-research0008.
- 1232 **Kourakis MJ**, Martindale MQ. Combined-method phylogenetic analysis of Hox and ParaHox genes of the Meta-
1233 zoa. *J Exp Zool*. 2000 Aug; 288(2):175–191.
- 1234 **Krishnan A**, Almén MS, Fredriksson R, Schiöth HB. The origin of GPCRs: identification of mammalian like
1235 Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi. *PLoS One*. 2012; 7:e29817. doi: 10.1371/jour-
1236 [nal.pone.0029817](https://doi.org/10.1371/journal.pone.0029817).
- 1237 **Kriventseva EV**, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM. Or-
1238 thoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res*.
1239 2015 Jan; 43(Database issue):D250–D256. <http://dx.doi.org/10.1093/nar/gku1220>, doi: 10.1093/nar/gku1220.
- 1240 **Kunst M**, Hughes ME, Raccuglia D, Felix M, Li M, Barnett G, Duah J, Nitabach MN. Calcitonin gene-related peptide
1241 neurons mediate sleep-specific circadian output in *Drosophila*. *Current Biology*. 2014 Nov; 24:2652–2664.
1242 doi: 10.1016/j.cub.2014.09.077.
- 1243 **Kyrchanova O**, Zolotarev N, Mogila V, Maksimenko O, Schedl P, Georgiev P. Architectural protein Pita coop-
1244 erates with dCTCF in organization of functional boundaries in Bithorax complex. *Development*. 2017 Jul;
1245 144:2663–2672. doi: 10.1242/dev.149815.
- 1246 **Ladoukakis E**, Pereira V, Magny EG, Eyre-Walker A, Couso JP. Hundreds of putatively functional small open
1247 reading frames in *Drosophila*. *Genome Biol*. 2011; 12(11):R118. <http://dx.doi.org/10.1186/gb-2011-12-11-r118>,
1248 doi: 10.1186/gb-2011-12-11-r118.
- 1249 **Larroux C**, Fahey B, Degnan SM, Adamski M, Rokhsar DS, Degnan BM. The NK homeobox gene cluster predates
1250 the origin of Hox genes. *Curr Biol*. 2007; 17(8):706–10. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17379523)
1251 [cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17379523](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17379523).
- 1252 **Laumer CE**, Bekkouche N, Kerbl A, Goetz F, Neves RC, Sørensen MV, Kristensen RM, Hejnlø A, Dunn CW, Giri-
1253 bet G, Worsaae K. Spiralian phylogeny informs the evolution of microscopic lineages. *Curr Biol*. 2015 Aug;
1254 25:2000–2006. doi: 10.1016/j.cub.2015.06.068.
- 1255 **Lauri A**, Bertucci P, Arendt D. Neurotrophin, p75, and Trk Signaling Module in the Developing Nervous Sys-
1256 tem of the Marine Annelid *Platynereis dumerilii*. *BioMed Research International*. 2016; 2016:2456062. doi:
1257 10.1155/2016/2456062.
- 1258 **Li G**, Liu X, Xing C, Zhang H, Shimeld SM, Wang Y. Cerberus-Nodal-Lefty-Pitx signaling cascade con-
1259 trols left-right asymmetry in *Amphioxus*. *Proc Natl Acad Sci U S A*. 2017 Apr; 114:3684–3689. doi:
1260 10.1073/pnas.1620519114.
- 1261 **Li L**, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*.
1262 2003; 13(9):2178–89. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=12952885)
1263 [retmode=ref&id=12952885](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=12952885).
- 1264 **Li W**, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide se-
1265 quences. *Bioinformatics*. 2006 Jul; 22(13):1658–1659. <http://dx.doi.org/10.1093/bioinformatics/btl158>, doi:
1266 10.1093/bioinformatics/btl158.
- 1267 **Li Y**, Cheng CN, Verdun VA, Wingert RA. Zebrafish nephrogenesis is regulated by interactions between retinoic
1268 acid, mecom, and Notch signaling. *Dev Biol*. 2014 Feb; 386:111–122. doi: 10.1016/j.ydbio.2013.11.021.
- 1269 **Liguz-Lecznar M**, Urban-Ciecko J, Kossut M. Somatostatin and Somatostatin-Containing Neurons in Shaping
1270 Neuronal Activity and Plasticity. *Frontiers in Neural Circuits*. 2016; 10:48. doi: 10.3389/fncir.2016.00048.
- 1271 **Lindemans M**, Janssen T, Husson SJ, Meelkop E, Temmerman L, Clynen E, Mertens I, Schoofs L. A neuromedin-
1272 pyrokinin-like neuropeptide signaling system in *Caenorhabditis elegans*. *Biochemical and Biophysical Re-*
1273 *search Communications*. 2009 Feb; 379:760–764. doi: 10.1016/j.bbrc.2008.12.121.
- 1274 **Lopez P**, Casane D, Philippe H. Heterotachy, an important process of protein evolution. *Molecular Biology and*
1275 *Evolution*. 2002 Jan; 19:1–7. doi: 10.1093/oxfordjournals.molbev.a003973.
- 1276 **Lowery LA**, Van Vactor D. The trip of the tip: understanding the growth cone machinery. *Nat Rev Mol Cell Biol*.
1277 2009 May; 10:332–343. doi: 10.1038/nrm2679.

- 1278 **Lu B**, Pang PT, Woo NH. The yin and yang of neurotrophin action. *Nat Rev Neurosci*. 2005 Aug; 6:603–614. doi:
1279 10.1038/nnr1726.
- 1280 **Luis Villanueva-Cañas J**, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Albà MM. New Genes and Functional
1281 Innovation in Mammals. *Genome Biology and Evolution*. 2017 Jul; 9:1886–1900. doi: 10.1093/gbe/evx136.
- 1282 **Luke GN**, Castro LFC, McLay K, Bird C, Coulson A, Holland PWH. Dispersal of NK homeobox gene clusters in
1283 *Amphioxus* and humans. *Proc Natl Acad Sci U S A*. 2003 Apr; 100(9):5292–5295. [http://dx.doi.org/10.1073/](http://dx.doi.org/10.1073/pnas.0836141100)
1284 [pnas.0836141100](http://dx.doi.org/10.1073/pnas.0836141100), doi: 10.1073/pnas.0836141100.
- 1285 **Mackowiak SD**, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuni G, Rajewsky N, Kempa S, Selbach
1286 M, Obermayer B. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol*.
1287 2015; 16:179. <http://dx.doi.org/10.1186/s13059-015-0742-x>, doi: 10.1186/s13059-015-0742-x.
- 1288 **Marshall CR**. EXPLAINING THE CAMBRIAN "EXPLOSION" OF ANIMALS. *Annu Rev Earth Planet Sci*.
1289 2006; 34(1):355–384. <http://www.annualreviews.org/doi/abs/10.1146/annurev.earth.33.031504.103001>, doi:
1290 10.1146/annurev.earth.33.031504.103001.
- 1291 **Martin JF**, Bradley A, Olson EN. The paired-like homeo box gene *MHox* is required for early events of skeleto-
1292 genesis in multiple lineages. *Genes Dev*. 1995 May; 9:1237–1249.
- 1293 **Martín-Durán JM**, Pang K, Børve A, Lê HS, Furu A, Cannon JT, Jondelius U, Hejnol A. Convergent evolution of
1294 bilaterian nerve cords. *Nature*. 2018 Jan; 553:45–50. doi: 10.1038/nature25030.
- 1295 **Maruyama YK**. A Sea Cucumber Homolog of the Mouse T-Brain-1 is Expressed in the Invaginated Cells of the
1296 Early Gastrula in *Holothuria leucospilota*. *Zoological Science*. 2000 Apr; 17:383–387. doi: 10.2108/jsz.17.383.
- 1297 **Mattar P**, Stevanovic M, Nad I, Cayouette M. *Cas21* controls higher-order nuclear organization in rod photore-
1298 ceptors. *Proc Natl Acad Sci U S A*. 2018 Aug; 115:E7987–E7996. doi: 10.1073/pnas.1803069115.
- 1299 **Matus DQ**, Thomsen GH, Martindale MQ. FGF signaling in gastrulation and neural development in *Ne-*
1300 *matostella vectensis*, an anthozoan cnidarian. *Dev Genes Evol*. 2007 Feb; 217:137–148. doi: 10.1007/s00427-
1301 006-0122-3.
- 1302 **Mayorova TD**, Kosevich IA. Serotonin-immunoreactive neural system and contractile system in the hydroid
1303 *Cladonema* (Cnidaria, Hydrozoa). *Invertebrate Neuroscience*. 2013 Dec; 13:99–106. doi: 10.1007/s10158-
1304 013-0152-2.
- 1305 **Mead PE**, Parganas E, Ohtsuka S, Morishita K, Gamer L, Kuliyeve E, Wright CVE, Ihle JN. *Evi-1* expression in
1306 *Xenopus*. *Gene Expression Patterns*. 2005 Jun; 5:601–608. doi: 10.1016/j.modgep.2005.03.007.
- 1307 **Melcher C**, Bader R, Walther S, Simakov O, Pankratz MJ. Neuromedin U and its putative *Drosophila* homolog
1308 *hugin*. *PLoS Biology*. 2006 Mar; 4:e68. doi: 10.1371/journal.pbio.0040068.
- 1309 **Michelson AM**, Abmayr SM, Bate M, Arias AM, Maniatis T. Expression of a *MyoD* family member prefigures
1310 muscle pattern in *Drosophila* embryos. *Genes Dev*. 1990 Dec; 4:2086–2097.
- 1311 **Milde S**, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TCG. Characterization of taxonomically
1312 restricted genes in a phylum-restricted cell type. *Genome Biol*. 2009; 10:R8. doi: 10.1186/gb-2009-10-1-r8.
- 1313 **Minh BQ**, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology*
1314 *and Evolution*. 2013 May; 30:1188–1195. doi: 10.1093/molbev/mst024.
- 1315 **Mirabeau O**, Joly JS. Molecular evolution of peptidergic signaling systems in bilaterians. *Proc Natl Acad Sci U S*
1316 *A*. 2013 May; 110:E2028–E2037. doi: 10.1073/pnas.1219956110.
- 1317 **Mita K**, Fujiwara S. Nodal regulates neural tube formation in the *Ciona intestinalis* embryo. *Dev Genes Evol*.
1318 2007 Aug; 217:593–601. doi: 10.1007/s00427-007-0168-x.
- 1319 **Mohan M**, Bartkuhn M, Herold M, Philippen A, Heintz N, Bardenhagen I, Leers J, White RA, Renkawitz-Pohl R,
1320 Saumweber H, Renkawitz R. The *Drosophila* insulator proteins CTCF and CP190 link enhancer blocking to
1321 body patterning. *EMBO J*. 2007; 26(19):4203–14. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17805343)
1322 [prlinks&dbfrom=pubmed&retmode=ref&id=17805343](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17805343).
- 1323 **Monteiro AS**, Schierwater B, Dellaporta SL, Holland PWH. A low diversity of ANTP class homeobox genes in
1324 Placozoa. *Evolution & Development*. 2006; 8:174–182. doi: 10.1111/j.1525-142X.2006.00087.x.

- 1325 **Moody WJ**, Simoncini L, Coombs JL, Spruce AE, Villaz M. Development of ion channels in early embryos. *Devel-*
1326 *opmental Neurobiology*. 1991; 22(7):674–684.
- 1327 **Moroz LL**, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov
1328 E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov
1329 YV, Swore JJ, et al. The ctenophore genome and the evolutionary origins of neural systems. *Nature*. 2014 Jun;
1330 510(7503):109–114. <http://dx.doi.org/10.1038/nature13400>, doi: 10.1038/nature13400.
- 1331 **Narendra V**, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, Reinberg D. CTCF establishes discrete functional
1332 chromatin domains at the Hox clusters during differentiation. *Science*. 2015 Feb; 347(6225):1017–1021. <http://dx.doi.org/10.1126/science.1262088>, doi: 10.1126/science.1262088.
- 1334 **Nguyen LT**, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for es-
1335 timating maximum-likelihood phylogenies. *Molecular Biology and Evolution*. 2015 Jan; 32:268–274. doi:
1336 10.1093/molbev/msu300.
- 1337 **Ocampo ID**, Zárate-Potes A, Pizarro V, Rojas CA, Vera NE, Cadavid LF. The immunotranscriptome of the
1338 Caribbean reef-building coral *Pseudodiploria strigosa*. *Immunogenetics*. 2015 Sep; 67:515–530. doi:
1339 10.1007/s00251-015-0854-1.
- 1340 **Okonechnikov K**, Golosova O, Fursov M, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit.
1341 *Bioinformatics*. 2012 Apr; 28:1166–1167. doi: 10.1093/bioinformatics/bts091.
- 1342 **Pai VP**, Willocq V, Pitcairn EJ, Lemire JM, Paré JF, Shi NQ, McLaughlin KA, Levin M. HCN4 ion channel function is
1343 required for early events that regulate anatomical left-right patterning in a nodal and lefty asymmetric gene
1344 expression-independent manner. *Biology Open*. 2017 Oct; 6:1445–1457. doi: 10.1242/bio.025957.
- 1345 **Panfilio KA**, Vargas Jentzsch IM, Benoit JB, Erezylmaz D, Suzuki Y, Colella S, Robertson HM, Poelchau MF, Wa-
1346 terhouse RM, Ioannidis P, Weirauch MT, Hughes DST, Murali SC, Werren JH, Jacobs CGC, Duncan EJ, Armisen
1347 D, Vreede BMI, Baa-Puyoulet P, Berger CS, et al. Molecular evolutionary trends and feeding ecology diversi-
1348 fication in the Hemiptera, anchored by the milkweed bug genome. *Genome Biology*. 2019 Apr; 20:64. doi:
1349 10.1186/s13059-019-1660-0.
- 1350 **Paps J**, Holland PWH. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty.
1351 *Nature Communications*. 2018 Apr; 9:1730. doi: 10.1038/s41467-018-04136-5.
- 1352 **Parra G**, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioin-*
1353 *formatics*. 2007 May; 23(9):1061–1067. <http://dx.doi.org/10.1093/bioinformatics/btm071>, doi: 10.1093/bioin-
1354 formatics/btm071.
- 1355 **Peterson KJ**, Dietrich MR, McPeck MA. MicroRNAs and metazoan macroevolution: insights into canalization,
1356 complexity, and the Cambrian explosion. *Bioessays*. 2009; 31(7):736–47. [http://eutils.ncbi.nlm.nih.gov/entrez/](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19472371)
1357 [eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19472371](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19472371).
- 1358 **Petryszak R**, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, Füllgrabe A, Fuentes AMP, Jupp S, Koskinen S,
1359 Mannion O, Huerta L, Megy K, Snow C, Williams E, Barzine M, Hastings E, Weissner H, Wright J, Jaiswal P, et al.
1360 Expression Atlas update—an integrated database of gene and protein expression in humans, animals and
1361 plants. *Nucleic Acids Res*. 2016 Jan; 44:D746–D752. doi: 10.1093/nar/gkv1045.
- 1362 **Philippe H**, Lartillot N, Brinkmann H. Multigene analyses of bilaterian animals corroborate the monophyly of
1363 Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol*. 2005; 22(5):1246–53. [http://eutils.ncbi.nlm.nih.](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15703236)
1364 [gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15703236](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15703236).
- 1365 **Phillips-Cremens JE**, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong CT, Hookway TA, Guo C, Sun
1366 Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG. Architectural protein
1367 subclasses shape 3D organization of genomes during lineage commitment. *Cell*. 2013; 153(6):1281–95. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23706625>.
- 1368 <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23706625>.
- 1369 **Pisani D**, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G. Genomic data
1370 do not support comb jellies as the sister group to all other animals. *Proc Natl Acad Sci U S A*. 2015 Dec;
1371 112:15402–15407. doi: 10.1073/pnas.1518127112.
- 1372 **Prochnik SE**, Rokhsar DS, Aboobaker AA. Evidence for a microRNA expansion in the bilaterian ancestor. *Dev*
1373 *Genes Evol*. 2007; 217(1):73–7. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17103184)
1374 [pubmed&retmode=ref&id=17103184](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eflink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17103184).

- 1375 **Prud'homme B**, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB. Repeated morpho-
1376 logical evolution through cis-regulatory changes in a pleiotropic gene. *Nature*. 2006; 440(7087):1050–3. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16625197>.
1377
- 1378 **Punta M**, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J,
1379 Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database.
1380 *Nucleic Acids Res*. 2012 Jan; 40(Database issue):D290–D301. <http://dx.doi.org/10.1093/nar/gkr1065>, doi:
1381 10.1093/nar/gkr1065.
- 1382 **Putnam NH**, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov
1383 VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar
1384 DS. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*.
1385 2007; 317(5834):86–94. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17615350>.
1386
- 1387 **Raj A**, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, Stephens M, Gilad Y, Pritchard JK. Thousands of novel
1388 translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*. 2016; 5. <http://dx.doi.org/10.7554/eLife.13328>, doi: 10.7554/eLife.13328.
1389
- 1390 **Ravisankar V**, Singh TP, Manoj N. Molecular evolution of the EGF-CFC protein family. *Gene*. 2011 Aug; 482:43–
1391 50. doi: 10.1016/j.gene.2011.05.007.
- 1392 **Rice P**, Longden I, Bleasby A. EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet*.
1393 2000; 16(6):276–7. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=10827456>.
1394
- 1395 **Richter DJ**, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem
1396 lineage. *eLife*. 2018 May; 7. doi: 10.7554/eLife.34226.
- 1397 **Rothberg JM**, Jacobs JR, Goodman CS, Artavanis-Tsakonas S. Slit: an extracellular protein necessary for devel-
1398 opment of midline glia and commissural axon pathways contains both EGF and LRR domains. *Genes Dev*.
1399 1990 Dec; 4:2169–2187.
- 1400 **Rousseau M**, Crutchley JL, Miura H, Suderman M, Blanchette M, Dostie J. Hox in motion: tracking HoxA cluster
1401 conformation during differentiation. *Nucleic Acids Res*. 2014 Feb; 42(3):1524–1540. <http://dx.doi.org/10.1093/nar/gkt998>, doi: 10.1093/nar/gkt998.
1402
- 1403 **Ryan JF**, Burton PM, Mazza ME, Kwong GK, Mullikin JC, Finnerty JR. The cnidarian-bilaterian ancestor possessed
1404 at least 56 homeoboxes: evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biol*. 2006;
1405 7(7):R64. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=16867185>.
1406
- 1407 **Ryan JF**, Mazza ME, Pang K, Matus DQ, Baxevas AD, Martindale MQ, Finnerty JR. Pre-bilaterian origins of the
1408 Hox cluster and the Hox code: evidence from the sea anemone, *Nematostella vectensis*. *PLoS One*. 2007;
1409 2(1):e153. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=17252055>.
1410
- 1411 **Ryan JF**, Pang K, Schnitzler CE, Nguyen AD, Moreland RT, Simmons DK, Koch BJ, Francis WR, Havlak P, NISCCSP,
1412 Smith SA, Putnam NH, Haddock SHD, Dunn CW, Wolfsberg TG, Mullikin JC, Martindale MQ, Baxevas AD. The
1413 genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science*. 2013 Dec;
1414 342(6164):1242592. <http://dx.doi.org/10.1126/science.1242592>, doi: 10.1126/science.1242592.
- 1415 **Ryan K**, Garrett N, Mitchell A, Gurdon JB. Eomesodermin, a key early gene in *Xenopus* mesoderm differentiation.
1416 *Cell*. 1996 Dec; 87:989–1000.
- 1417 **Röttinger E**, DuBuc TQ, Amiel AR, Martindale MQ. Nodal signaling is required for mesodermal and ventral
1418 but not for dorsal fates in the indirect developing hemichordate, *Ptychodera flava*. *Biology Open*. 2015 May;
1419 4:830–842. doi: 10.1242/bio.011809.
- 1420 **Santos ME**, Le Bouquin A, Crumière AJJ, Khila A. Taxon-restricted genes at the origin of a novel trait allowing
1421 access to a new environment. *Science*. 2017 Oct; 358:386–390. doi: 10.1126/science.aan2748.
- 1422 **Saudemont A**, Dray N, Hudry B, Le Gouar M, Vervoort M, Balavoine G. Complementary striped expression
1423 patterns of NK homeobox genes during segment formation in the annelid *Platynereis*. *Dev Biol*. 2008 May;
1424 317:430–443. doi: 10.1016/j.ydbio.2008.02.013.

- 1425 **Schwaiger M**, Schönaauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF, Schinko JB, Renfer E, Fredman D,
1426 Technau U. Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res.* 2014
1427 Apr; 24:639–650. doi: [10.1101/gr.162529.113](https://doi.org/10.1101/gr.162529.113).
- 1428 **Sebé-Pedrós A**, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, Adamski M, Adamska M, Hughes
1429 TR, Gómez-Skarmeta JL, Ruiz-Trillo I. Early evolution of the T-box transcription factor family. *Proc Natl Acad*
1430 *Sci U S A.* 2013 Oct; 110:16050–16055. doi: [10.1073/pnas.1309748110](https://doi.org/10.1073/pnas.1309748110).
- 1431 **Sharma V**, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M. A genomics approach reveals insights
1432 into the importance of gene losses for mammalian adaptations. *Nat Commun.* 2018 Mar; 9:1215. doi:
1433 [10.1038/s41467-018-03667-1](https://doi.org/10.1038/s41467-018-03667-1).
- 1434 **Shen MM**. Nodal signaling: developmental roles and regulation. *Development.* 2007 Mar; 134:1023–1034. doi:
1435 [10.1242/dev.000166](https://doi.org/10.1242/dev.000166).
- 1436 **Simakov O**, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D,
1437 Savage R, Osoegawa K, de Jong P, Grimwood J, Chapman JA, Shapiro H, Aerts A, Otillar RP, Terry AY, Boore JL,
1438 et al. Insights into bilaterian evolution from three spiralian genomes. *Nature.* 2013; 493(7433):526–31. [http:](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23254933)
1439 [//eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23254933](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=23254933).
- 1440 **Simion P**, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Queinnec E, Ereskovsky
1441 A, Lapebie P, Corre E, Delsuc F, King N, Worheide G, Manuel M. A Large and Consistent Phylogenomic
1442 Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr Biol.* 2017 Apr; 27:958–967. doi:
1443 [10.1016/j.cub.2017.02.031](https://doi.org/10.1016/j.cub.2017.02.031).
- 1444 **Soding J**. Protein homology detection by HMM-HMM comparison. *Bioinformatics.* 2005; 21(7):951–60. [http:](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15531603)
1445 [//eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15531603](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=15531603).
- 1446 **Som A**. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics.* 2014
1447 May; 16:536–548. doi: [10.1093/bib/bbu015](https://doi.org/10.1093/bib/bbu015).
- 1448 **Srivastava M**, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Car-
1449 penter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss
1450 LW, Schierwater B, Dellaporta SL, et al. The Trichoplax genome and the nature of placozoans. *Nature.*
1451 2008; 454(7207):955–60. [http:](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18719581)
1452 [//eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18719581](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18719581).
- 1453 **Srivastava M**, Simakov O, Chapman J, Fahey B, Gauthier ME, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten
1454 U, Larroux C, Putnam NH, Stanke M, Adamska M, Darling A, Degnan SM, Oakley TH, Plachetzki DC, Zhai Y,
1455 Adamski M, et al. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature.*
1456 2010; 466(7307):720–6. [http:](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=20686567)
1457 [//eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=20686567](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=20686567).
- 1458 **St Pierre SE**, Galindo MI, Couso JP, Thor S. Control of *Drosophila* imaginal disc development by rotund and
1459 roughened eye: differentially expressed transcripts of the same gene encoding functionally distinct zinc
1460 finger proteins. *Development.* 2002 Mar; 129:1273–1281.
- 1461 **Steinke D**, Salzburger W, Braasch I, Meyer A. Many genes in fish have species-specific asymmetric rates of
1462 molecular evolution. *BMC Genomics.* 2006 Feb; 7:20. doi: [10.1186/1471-2164-7-20](https://doi.org/10.1186/1471-2164-7-20).
- 1463 **Stergiopoulos A**, Elkouris M, Politis PK. Prospero-related homeobox 1 (Prox1) at the crossroads of diverse path-
1464 ways during adult neural fate specification. *Frontiers in Cellular Neuroscience.* 2014; 8:454. doi: [10.3389/fn-](https://doi.org/10.3389/fn-cel.2014.00454)
1465 [cel.2014.00454](https://doi.org/10.3389/fn-cel.2014.00454).
- 1466 **Su YH**, Yu JK. EvoDevo: Changes in developmental controls underlying the evolution of animal body plans.
1467 *Developmental Biology.* 2017 Jul; 427:177–178. doi: [10.1016/j.ydbio.2017.05.023](https://doi.org/10.1016/j.ydbio.2017.05.023).
- 1468 **Sun X**, Lin Y. Npas4: Linking Neuronal Activity to Memory. *Trends in Neurosciences.* 2016 Apr; 39:264–275. doi:
1469 [10.1016/j.tins.2016.02.003](https://doi.org/10.1016/j.tins.2016.02.003).
- 1470 **Suo S**, Ishiura S, Van Tol HHM. Dopamine receptors in *C. elegans*. *European Journal of Pharmacology.* 2004
1471 Oct; 500:159–166. doi: [10.1016/j.ejphar.2004.07.021](https://doi.org/10.1016/j.ejphar.2004.07.021).
- 1472 **Szklarczyk D**, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen
1473 LJ, von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks,
1474 made broadly accessible. *Nucleic Acids Res.* 2017 Jan; 45:D362–D368. doi: [10.1093/nar/gkw937](https://doi.org/10.1093/nar/gkw937).

- 1475 **Ségalat L**, Elkes DA, Kaplan JM. Modulation of serotonin-controlled behaviors by Go in *Caenorhabditis elegans*.
1476 *Science*. 1995 Mar; 267:1648–1651. doi: [10.1126/science.7886454](https://doi.org/10.1126/science.7886454).
- 1477 **Tagawa K**, Humphreys T, Satoh N. T-Brain expression in the apical organ of hemichordate tornaria larvae
1478 suggests its evolutionary link to the vertebrate forebrain. *J Exp Zool*. 2000 Apr; 288:23–31.
- 1479 **Tapscott SJ**, Davis RL, Thayer MJ, Cheng PF, Weintraub H, Lassar AB. MyoD1: a nuclear phosphoprotein requir-
1480 ing a Myc homology region to convert fibroblasts to myoblasts. *Science*. 1988 Oct; 242:405–411.
- 1481 **Tatusov RL**, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science*. 1997; 278(5338):631–
1482 7. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=9381173)
1483 [9381173](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=9381173).
- 1484 **Technau U**, Rudd S, Maxwell P, Gordon PMK, Saina M, Grasso LC, Hayward DC, Sensen CW, Saint R, Holstein
1485 TW, Ball EE, Miller DJ. Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians.
1486 *Trends Genet*. 2005 Dec; 21:633–639. doi: [10.1016/j.tig.2005.09.007](https://doi.org/10.1016/j.tig.2005.09.007).
- 1487 **Teixeira CM**, Rosen ZB, Suri D, Sun Q, Hersh M, Sargin D, Dincheva I, Morgan AA, Spivack S, Krok AC, Hirschfeld-
1488 Stoler T, Lambe EK, Siegelbaum SA, Ansorge MS. Hippocampal 5-HT Input Regulates Memory Formation and
1489 Schaffer Collateral Excitation. *Neuron*. 2018 Jun; 98(5):992–1004. doi: [10.1016/j.neuron.2018.04.030](https://doi.org/10.1016/j.neuron.2018.04.030).
- 1490 **Tekaia F**. Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights*. 2016; 9:17–28. doi:
1491 [10.4137/GEI.S37925](https://doi.org/10.4137/GEI.S37925).
- 1492 **Telford MJ**, Budd GE, Philippe H. Phylogenomic Insights into Animal Evolution. *Curr Biol*. 2015 Oct; 25:R876–
1493 R887. doi: [10.1016/j.cub.2015.07.060](https://doi.org/10.1016/j.cub.2015.07.060).
- 1494 **Tena JJ**, Neto A, de la Calle-Mustienes E, Bras-Pereira C, Casares F, Gómez-Skarmeta JL. Odd-skipped
1495 genes encode repressors that control kidney development. *Dev Biol*. 2007 Jan; 301:518–531. doi:
1496 [10.1016/j.ydbio.2006.08.063](https://doi.org/10.1016/j.ydbio.2006.08.063).
- 1497 **Thomas GWC**, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA, Ayoub NA, Batterham
1498 P, Bellair M, Binford GJ, Chao H, Chen YH, Childers C, Dinh H, Doddapaneni HV, Duan JJ, Dugan S, Esposito
1499 LA, Friedrich M, et al. Gene content evolution in the arthropods. *Genome Biology*. 2020 Jan; 21:15. doi:
1500 [10.1186/s13059-019-1925-7](https://doi.org/10.1186/s13059-019-1925-7).
- 1501 **Thomas-Chollier M**, Ledent V, Leyns L, Vervoort M. A non-tree-based comprehensive study of metazoan Hox
1502 and ParaHox genes prompts new insights into their origin and evolution. *BMC Evol Biol*. 2010; 10:73. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=20222951>.
1503
- 1504 **Tomancak P**, Beaton A, Weiszmam R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker
1505 SE, Rubin GM. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis.
1506 *Genome Biol*. 2002; 3:RESEARCH0088.
- 1507 **Torruella G**, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, del Campo J, Eme L, Pérez-Cordón G, Whipps CM,
1508 Nichols KM, Paley R, Roger AJ, Sitjà-Bobadilla A, Donachie S, Ruiz-Trillo I. Phylogenomics Reveals Convergent
1509 Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Curr Biol*. 2015 Sep; 25:2404–2410. doi:
1510 [10.1016/j.cub.2015.07.053](https://doi.org/10.1016/j.cub.2015.07.053).
- 1511 **Trachana K**, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. Orthology prediction methods: a quality
1512 assessment using curated protein families. *Bioessays*. 2011; 33(10):769–80. [http://eutils.ncbi.nlm.nih.gov/](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=21853451)
1513 [entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=21853451](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=21853451).
- 1514 **Ueno T**, Tomita J, Tanimoto H, Endo K, Ito K, Kume S, Kume K. Identification of a dopamine pathway that regu-
1515 lates sleep and arousal in *Drosophila*. *Nature Neuroscience*. 2012 Nov; 15:1516–1523. doi: [10.1038/nn.3238](https://doi.org/10.1038/nn.3238).
- 1516 **Van Bortle K**, Ramos E, Takenaka N, Yang J, Wahi JE, Corces VG. *Drosophila* CTCF tandemly aligns with other
1517 insulator proteins at the borders of H3K27me3 domains. *Genome Res*. 2012; 22(11):2176–87. [http://eutils.](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=22722341)
1518 [ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=22722341](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=22722341).
- 1519 **Vidal-Gadea A**, Topper S, Young L, Crisp A, Kressin L, Elbel E, Maples T, Brauner M, Erbguth K, Axelrod A,
1520 Gottschalk A, Siegel D, Pierce-Shimomura JT. *Caenorhabditis elegans* selects distinct crawling and swimming
1521 gaits via dopamine and serotonin. *Proceedings of the National Academy of Sciences of the United States of*
1522 *America*. 2011 Oct; 108:17504–17509. doi: [10.1073/pnas.1108673108](https://doi.org/10.1073/pnas.1108673108).
- 1523 **Vietri Rudan M**, Hadjur S. Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution. *Trends*
1524 *Genet*. 2015 Nov; 31(11):651–660. <http://dx.doi.org/10.1016/j.tig.2015.09.004>, doi: [10.1016/j.tig.2015.09.004](https://doi.org/10.1016/j.tig.2015.09.004).

- 1525 **Wang Q**, Lan Y, Cho ES, Maltby KM, Jiang R. Odd-skipped related 1 (Odd 1) is an essential regulator of heart
1526 and urogenital development. *Dev Biol.* 2005 Dec; 288:582–594. doi: [10.1016/j.ydbio.2005.09.024](https://doi.org/10.1016/j.ydbio.2005.09.024).
- 1527 **Ward N**, Moreno-Hagelsieb G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST:
1528 how much do we miss? *PloS One.* 2014; 9:e101850. doi: [10.1371/journal.pone.0101850](https://doi.org/10.1371/journal.pone.0101850).
- 1529 **Watanabe H**, Schmidt HA, Kuhn A, Höger SK, Kocagöz Y, Laumann-Lipp N, Ozbek S, Holstein TW. Nodal sig-
1530 nalling determines biradial asymmetry in Hydra. *Nature.* 2014 Nov; 515(7525):112–115. [http://dx.doi.org/10.](http://dx.doi.org/10.1038/nature13666)
1531 [1038/nature13666](http://dx.doi.org/10.1038/nature13666), doi: 10.1038/nature13666.
- 1532 **Weintraub AS**, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley
1533 DL, Guo YE, Hnisz D, Jaenisch R, Bradner JE, Gray NS, Young RA. YY1 Is a Structural Regulator of Enhancer-
1534 Promoter Loops. *Cell.* 2017 Dec; 171:1573–1588.e28. doi: [10.1016/j.cell.2017.11.008](https://doi.org/10.1016/j.cell.2017.11.008).
- 1535 **Weisberg E**, Winnier GE, Chen X, Farnsworth CL, Hogan BL, Whitman M. A mouse homologue of FAST-1 trans-
1536 duces TGF β superfamily signals and is expressed during early embryogenesis. *Mechanisms of Development.*
1537 1998 Dec; 79:17–27.
- 1538 **Weisman CM**, Murray AW, Eddy SR. Many but not all lineage-specific genes can be explained by homology
1539 detection failure. *bioRxiv.* 2020; <https://www.biorxiv.org/content/early/2020/02/28/2020.02.27.968420>, doi:
1540 [10.1101/2020.02.27.968420](https://doi.org/10.1101/2020.02.27.968420).
- 1541 **Wheeler BM**, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ. The deep evolution of
1542 metazoan microRNAs. *Evol Dev.* 2009; 11(1):50–68. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19196333)
1543 [cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19196333](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=19196333).
- 1544 **Whelan NV**, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. Ctenophore
1545 relationships and their placement as the sister group to all other animals. *Nature Ecology & Evolution.* 2017
1546 Nov; 1:1737–1746. doi: 10.1038/s41559-017-0331-3.
- 1547 **Wilson KHS**. The genome sequence of the protostome *Daphnia pulex* encodes respective orthologues of a
1548 neurotrophin, a Trk and a p75NTR: evolution of neurotrophin signaling components and related proteins in
1549 the Bilateria. *BMC Evol Biol.* 2009 Oct; 9:243. doi: 10.1186/1471-2148-9-243.
- 1550 **de Wit J**, Hong W, Luo L, Ghosh A. Role of leucine-rich repeat proteins in the development and function of
1551 neural circuits. *Annual Review of Cell and Developmental Biology.* 2011; 27:697–729. doi: 10.1146/annurev-
1552 cellbio-092910-154111.
- 1553 **Wu J**, Susko E. A test for heterotachy using multiple pairs of sequences. *Molecular Biology and Evolution.* 2011
1554 May; 28:1661–1673. doi: 10.1093/molbev/msq346.
- 1555 **Yan YT**, Gritsman K, Ding J, Burdine RD, Corrales JD, Price SM, Talbot WS, Schier AF, Shen MM. Conserved
1556 requirement for EGF-CFC genes in vertebrate left-right axis formation. *Genes Dev.* 1999 Oct; 13:2527–2537.
- 1557 **Yao LC**, Blitz IL, Peiffer DA, Phin S, Wang Y, Ogata S, Cho KKY, Arora K, Warrior R. Schnurri transcription factors
1558 from *Drosophila* and vertebrates can mediate Bmp signaling through a phylogenetically conserved mecha-
1559 nism. *Development.* 2006 Oct; 133:4025–4034. doi: [10.1242/dev.02561](https://doi.org/10.1242/dev.02561).
- 1560 **Yu JK**, Mazet F, Chen YT, Huang SW, Jung KC, Shimeld SM. The Fox genes of *Branchiostoma floridae*. *Dev Genes*
1561 *Evol.* 2008 Dec; 218:629–638. doi: 10.1007/s00427-008-0229-9.
- 1562 **Zemach A**, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methy-
1563 lation. *Science.* 2010 May; 328(5980):916–919. <http://dx.doi.org/10.1126/science.1186366>, doi: [10.1126/sci-](https://doi.org/10.1126/science.1186366)
1564 [ence.1186366](https://doi.org/10.1126/science.1186366).
- 1565 **Zhou S**, Zawel L, Lengauer C, Kinzler KW, Vogelstein B. Characterization of human FAST-1, a TGF β and activin
1566 signal transducer. *Molecular Cell.* 1998 Jul; 2:121–127.

Appendix 1

Orthology pipeline and clustering

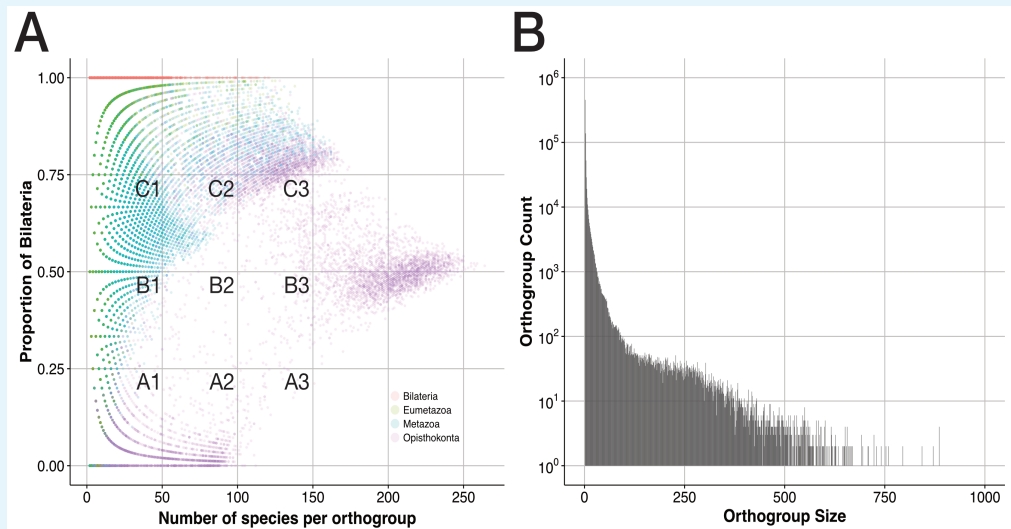
To generate clusters of orthologous proteins from the collected sequence data, we used the OrthoMCL pipeline (*Li et al., 2003*). OrthoMCL is a graph-based method for ortholog group identification that represents sequences as nodes and their similarities as weighted edges. A normalization step adjusts initial similarity scores to reflect species distance and ensures that edge weights for sequence pairs are comparable between different genomes. Finally, the Markov cluster algorithm (*van Dongen, 2000*) performs random walks on the normalized graph by simulating transition probabilities of sequences to other nodes, thereby revealing an underlying cluster structure. To create the BLAST similarity table required by OrthoMCL, we performed all-vs-all BLAST searches with 124 million sequences (with default BLAST parameters, except «-outfmt 6»; BLAST version 2.2.28). Roughly one million CPU hours were necessary for this task, running hundreds of jobs in parallel on a high performance computing platform. Merging the individual output files, we obtained a similarity score table of ~500 GB, containing roughly 6 billion BLAST hits (see Supplementary File 1–Supplementary Table 5). In the original implementation, OrthoMCL loads the BLAST output table into a MySQL database and performs subsequent computations within the relational database. Because of its size, we could not load the BLAST output table into a physical MySQL database. We therefore ported all MySQL processes to the statistical computing environment R to execute them in computer memory. Test experiments, carried out in parallel with our R implementation and the original software, produced identical results, demonstrating that the R version of OrthoMCL accurately reproduces the outcome of the standard pipeline (Supplementary File 1–Supplementary Table 4). After obtaining the final table with adjusted pairwise distance information in R, we used Markov clustering (*van Dongen, 2000*), as in the original protocol, to combine sequences to orthologous groups.

Depending on the origin of compared sequences, OrthoMCL creates three ortholog tables: a table with reciprocal relationships of sequences between different species (ortholog table), a table of within-species relationships (in-paralogs), and a table of co-orthologs with protein pairs that are connected through orthology and in-paralogy. Of 124 million gathered sequences, 122 million had at least one blast hit in the database, giving rise to a collection of 6 billion blast pairs as raw material for orthology clustering. The OrthoMCL pipeline retained 35 million of these sequences in 806 million pairs of the three orthology tables. Thus, 28.8 % of the sequences had enough similarity with other sequences to participate in orthology group construction whereas the majority of input sequences were so remotely related to other sequences that our pipeline could not merge them with a cluster. As expected, artificially generated ORFs represented by far the largest portion of the non-clustered sequences (91.3 %).

As we observed that a large in-paralog table (5.8× larger than the ortholog plus co-ortholog tables for the final dataset) negatively affected the accuracy of the clustering process, we omitted this table in subsequent trials. In the final MCL run, we obtained 824,605 orthologous groups with 6,743,519 distinct sequences derived from 118,499,524 protein pairs (blast hits) of the ortholog and co-ortholog tables. Discarding the large in-paralog table led to a drop in the percentage of clustered sequences from 28.8 % to 5.5 %, indicating that a considerable amount of orthogroups in the larger dataset consisted of paralogs (Supplementary File 1–Supplementary Table 5).

To investigate the properties of orthogroups as old as bilaterians or older, we plotted for the respective orthogroups the number of species against their proportion of bilaterians (Appendix 1–Figure 1). Position and abundance of many data points in the resulting plot are

a consequence of dataset composition. For example, (i) the majority of orthogroups is small, leading to an abundance of solid (because of overlap) data points for small orthogroups (Appendix 1–Figure 1, left part; Supplementary File 1–Supplementary Table 5); (ii) bilaterians and non-bilaterians including fungi are groups roughly equal in size (142 vs. 131 species), preventing that bilaterian sequences exceed a coverage of ~50 % in large orthogroups. Similarly, bilaterian content can hardly fall below 40 % to 50 % in large orthogroups with more than ~175 species, giving rise to an arrowhead shape at the right side of the plot (Appendix 1–Figure 1). (iii) orthogroups with a bilaterian ancestor have, by definition, a bilaterian content of 100 % and are therefore spread as dotted red line on top of the plot that is fading away in orthogroups with more than 100 species; (iv) orthogroups with metazoan and eumetazoan ancestor (green and blue) concentrate on the left part of the plot because not more than 33 non-bilaterian metazoans are present in the dataset, restricting orthogroup size. In addition, the low orthogroup density in sectors B2, B3, and C3 suggests that ancient genes, that evolved in the ancestor of eumetazoans or earlier and survived in bilaterians, do not get lost randomly at multiple nodes in the bilaterian tree. Instead, they tend to be maintained across most bilaterian species. It remains to be seen whether this behaviour is specific for bilaterians in this dataset or a general evolutionary pattern.



Appendix 1 Figure 1. General properties of sequence clusters from a bilaterian viewpoint. A:

The proportion of bilaterians per orthogroup is shown as a function of orthogroup size (in terms of species number) for 207,285 orthogroups that trace back to the four ancestors Bilateria, Eumetazoa, Metazoa, and Opisthokonta. Dot colours indicate the orthogroup ancestor and are printed with 85 % transparency to reveal overlaps. B: Orthogroup count (how often orthogroups of a given size are observed) is displayed as function of orthogroup size (number of sequences present in an orthogroup). 34 orthogroups with more than 1,000 sequences were omitted. Almost all of these sizes occurred only once.

Cluster evaluation and quality control

In a first approach to verify the accuracy of our clustering results, we employed as an external benchmark a manually curated gene set of 70 orthologous groups (Trachana et al., 2011), the orthobench dataset (<http://egglog.embl.de/orthobench>). For the members of every orthobench family, we determined the corresponding BigWenDB sequence ID and the cluster ID (orthogroup ID) to which this sequence was assigned during clustering. We then analysed how the members of a given orthobench family were distributed among orthogroups in the BigWenDB. We performed such comparisons for two MCL inflation parameters ($I =$

1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696

1.3 and $I = 1.4$) and two database sizes (full database and database without paralog table). The clustering with the highest agreement to the expected orthobench outcome was the dataset with inflation parameter $I = 1.3$ and without paralog table (mcl_ortho-coortho_1.3.7; see Supplementary File 3). In this dataset, 46 of 70 protein families were assigned correctly, i. e. in 65.7 % of the cases our pipeline combined all members of an orthobench family, as expected, in a single orthogroup. However, blast hits that allow correct mapping were not found for all orthobench family members, and some members were mapped to erroneously predicted proteins. In such cases, orthobench members may be linked to an orthogroup different from the rest of the family, leading to the impression that several orthogroups exist for this family. According to our estimates, such mapping errors reduce accuracy by at least 5 %, suggesting a correct orthology inference rate above 70 % for our dataset. In contrast, only 10 % to 48 % of reference orthogroups were predicted correctly in the orthobench comparison (Trachana et al., 2011), indicating that the representative coverage of our dataset considerably improves orthogroup inference quality.

Evolutionary relationships of homeodomain-containing genes are difficult to trace because of the strong conservation and shortness of the homeodomain (60 AA) (Irvine et al., 1997; Kourakis and Martindale, 2000). To understand how our study deals with these difficulties, we analysed the composition of orthogroups containing NK (Nirenberg-Kim) homeobox genes. Like Hox and ParaHox gene clusters, the NK cluster is a close association of homeobox genes with crucial roles in animal development. It consists of the six genes tinman, bagpipe, ladybird (early and late), C15, and slouch in *D. melanogaster*. They are all involved in mesodermal patterning (Kim and Nirenberg, 1989; Jagla et al., 2001). Genomic data from vertebrates and the cephalochordate *Branchiostoma* indicate that the NK cluster is an ancient feature of bilaterians, but has been duplicated and split repeatedly in chordate history, leading to the presence of four dispersed clusters and multiple paralogs of each gene in humans (Luke et al., 2003). Several rearrangements have also been observed in the NK cluster of arthropods (Chan et al., 2015). In addition, studies of the homeodomain gene complement of sponges and cnidarians revealed that NK cluster genes predate the evolution of bilaterians (Ryan et al., 2006; Larroux et al., 2007). Given these findings, we can expect that NK homeobox genes from diverse metazoans (sponges, cnidarians, vertebrates, and insects) are each represented in a single orthogroup. Analysing the orthogroups of all *Drosophila* and human NK cluster genes revealed that, indeed, bilaterian and non-bilaterian orthologs of the five NK genes were joined in five corresponding groups (Supplementary File 1–Supplementary Table 6). These five orthogroups contained sequences from 81–128 (of 142) bilaterian species, including the known *Drosophila* and human NK genes, as well as sponge, cnidarian, and ctenophore sequences. We found placozoan sequences in a single orthogroup, OG_613 (NKX2), suggesting the previously unknown existence of NK class homeobox genes in Placozoa (Monteiro et al., 2006). In contrast to other NK genes, *Drosophila* tinman is not located in the group of its vertebrate counterparts NKX2.3/2.5/2.6 (OG_613; Supplementary File 1–Supplementary Table 6). It has been shown previously that orthology relationships between tinman and vertebrate NKX2 genes are difficult to establish because of the fast evolving insect tinman genes (Harvey, 1996; Saudemont et al., 2008). In line with these observations, tinman was assigned to a small orthogroup restricted to endopterygote insects (OG_92160) while other putative NKX2 orthologs from a wide range of arthropods (32/37 species) were combined with vertebrate NKX2 genes in orthogroup OG_613.

Consistency between our method and an independent method would further underline the reliability of inferred orthogroups. We therefore prepared our data for a control run with the orthogroup inference algorithm OrthoFinder that, in contrast to OrthoMCL, takes into account a so far unrecognised gene length bias (Emms and Kelly, 2015). However,

the number of pairwise blast similarity tables, resembling OrthoFinder's input, increases quadratically with the number of species, and so does the amount of required main memory. With 80 species and 6,320 corresponding blast tables, approximately 250 GB of memory are occupied, precluding a run with the full dataset (273 species; 74,256 blast tables) on current computers. OrthoFinder thus cannot be used to confirm our data until it is adapted to large data sets, in turn illustrating the power of our modified version of the OrthoMCL pipeline.

Taken together, the assessment of clustering quality using a benchmark and a homeobox gene set indicates that orthology prediction in the BigWenDB accurately captures known evolutionary relationships of difficult target genes over large evolutionary distances. We conclude therefore that our cluster results are well suited as raw material for the search of bilaterian-specific genes.

Identification of bilaterian-specific genes

To infer lineage-specific genes, we determined on the basis of NCBI taxonomy (<ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz>) the last common ancestor of the species present in all 824,605 orthologous groups of the final clustering. Together with other ancient groups such as Metazoa, Eumetazoa, or Opisthokonta, the taxon Bilateria is among the top ten of taxa with the highest counts (42,946 bilaterian-specific orthogroups; Supplementary File 1–Supplementary Table 25). While these counts include all ortholog groups that trace back to a given ancestor, the majority of groups contains only few species (see **Figure 4**, Appendix 1–Figure 1, Supplementary File 1–Supplementary Table 5). To obtain meaningful groups with a broad representation across bilaterians, we required that at least 10 % of the species of each bilaterian super-phylum must be present (Ecdysozoa ≥ 6 , Lophotrochozoa ≥ 4 , and Deuterostomia ≥ 7 species). We included orthogroups with zero ecdysozoans or lophotrochozoans if the count for the two other super-phyla met the 10 % threshold, thereby allowing for the loss of bilaterian-specific genes in ecdysozoans or lophotrochozoans. Following these rationales, we obtained 345 bilaterian-specific groups.

At least four types of error might impair our set of bilaterian-specific orthologous groups: (1) An orthogroup is judged older than bilaterians, but is in fact bilaterian-specific (orthogroup too large), (2) an orthogroup is inferred to be bilaterian-specific, but is in fact older (orthogroup too small), (3) an orthogroup is found to be bilaterian-specific, but is in fact younger (orthogroup too large), (4) an orthogroup is considered younger than bilaterians, but is in fact bilaterian-specific (orthogroup too small).

The presence of several bilaterian sequences and a single sequence from an earlier branching eukaryote would conceal the potential bilaterian ancestry of an orthogroup (type 1 error). We therefore searched for ortholog groups with broad bilaterian representation, according to our above mentioned rules, and up to two outgroup sequences. Of 349 orthogroups satisfying these criteria, the majority (263 or 75.3 %) contained as outliers sequences of cnidarian origin, the sister group of bilaterians. To maximise the likelihood of detecting true outliers, we considered only organisms without direct sister group relationship for further analysis and obtained 86 additional bilaterian-specific candidate groups with one or two non-bilaterian/non-cnidarian sequences. As the probability is high that these orthogroups contain phylogenetically unrelated outliers and actually originated in the bilaterian ancestor, we ranked them, together with the 345 previous orthogroups, in a set of 431 bilaterian-specific orthogroups.

Type 2 errors can arise if the MCL algorithm does not combine a group with bilaterian ancestry and a group with related sequences from non-bilaterian species although both groups might represent a single natural orthology group. To identify such errors, we com-

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

1782

1783

1784

1785

1786

1787

1788

1789

1790

1791

1792

1793

1794

1795

1796

1797

puted for all 824,605 orthogroups multiple sequence alignments and turned them into profile hidden Markov models (HMMs) that describe alignment consensus sequences in a probabilistic way (Eddy, 1998). We then assembled a database from the HMMs and searched the two next similar profiles for every bilaterian-specific group using sensitive HMM-HMM alignments (Soding, 2005). We devised a new reciprocal HMM-HMM alignment comparison step, analogous to the strategy of reciprocal best blast hits (Tatusov et al., 1997; Ward and Moreno-Hagelsieb, 2014), to discover bidirectional best hit orthogroup pairs prognostic for common descent. To demonstrate the power of this method, we analysed the orthogroup distribution of two example proteins, Sprouty, an inhibitor of FGF signalling, and the insulator protein GAGA factor. We found that the orthogroups of both, *D. melanogaster* Sprouty and *D. melanogaster* GAGA factor, were smaller than anticipated considering their reported phylogenetic distribution (Matus et al., 2007; Heger et al., 2013). In both cases, the reciprocal best hit strategy enabled us to detect highly similar orthogroups with known Sprouty and GAGA factor orthologs that complemented the original orthogroup. After fusion of query and reciprocal best hit orthogroups, the resulting sequence collections matched the expected phylogenetic coverage (Supplementary File 1–Supplementary Table 7). Encouraged by these findings, we examined the 431 bilaterian-specific orthogroups accordingly and excluded orthogroups from the list if they satisfied three criteria: (i) their best or second best HMM-HMM database hit modifies the ancestor of the resulting fusion group, (ii) their best or second best hit orthogroup is a reciprocal best hit, and (iii) their best or second best hit orthogroup does not contain more than three bilaterian species. With the last criterion we avoid to eliminate orthogroups whose reciprocal best hit is an ancient orthogroup with wide bilaterian representation, an indicator of homology rather than of orthology. The majority of bilaterian-specific orthogroups (84.2 % or 363/431 orthogroups) were not affected by this procedure. Therefore we considered them high-confidence bilaterian-specific orthogroups. On the other hand, 68 bilaterian-specific orthogroups (15.8 %) were possibly false positives and may have originated in pre-bilaterian time.

If, for example, several insects and a single sequence from a vertebrate populate an orthogroup, a bilaterian ancestor would be computed for this group although, from a phylogenetic point of view, the single vertebrate sequence is more likely an outlier added to the group erroneously. The filtering rules mentioned above require that at least 10 % of the species in each super-phylum are present in a group to qualify as bilaterian-specific. They effectively prevent type 3 errors in our list of bilaterian-specific orthogroups that were caused by the addition of < 4 sequences. In contrast, we cannot currently prevent potentially wrong orthology inference if four or more sequences of an ancestor-changing lineage were added erroneously (e. g. four ecdysozoan sequences added to an otherwise mammalian-specific orthogroup). However, this error mainly affects small bilaterian-specific orthogroups with only few sequences from deuterostomes, lophotrochozoans, and/or ecdysozoans because of their lack in representativeness. Detailed phylogenetic analysis as well as improved taxon sampling would be necessary to discover such false-positive assignments.

Type 4 errors occur if an orthogroup is estimated younger than bilaterians, but is—accidentally—not joined with another, similar orthogroup that would convert the ancestor to Bilateria if combined with the original group (e. g. a vertebrate-specific orthogroup and a highly similar insect-specific orthogroup would create a bilaterian-specific orthogroup). To detect such errors, it is necessary to perform all-vs-all profile comparisons of the orthogroups younger than bilaterians. Next, combinations of similar groups need to be determined that would shift the former individual ancestors to a new common bilaterian ancestor and that are each other's bidirectional best hit. Due to the high computational investment we refrained from further investigating this error source in this manuscript.

To further probe accuracy of the 363 bilaterian-specific orthogroups, we mapped human

1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817

and *D. melanogaster* sequences contained in these orthogroups to the respective genome (versions hg38 and dm6) using BLAT (**Kent, 2002**). Such mapping was possible for 348/363 orthogroups (95.87 %). We then checked whether the target gene to which these sequences were assigned, belonged to the initial orthogroup. This was not true in a considerable number of cases. For example, often bilaterian-specific orthogroups contained short ORFs from *H. sapiens* or *D. melanogaster* that mapped to a particular gene. The corresponding full length protein, however, was assigned to a different orthogroup with a different ancestor, indicating that separation of genes into two or more orthogroups affected integrity of the 363 orthogroups set. We therefore excluded all orthogroups with potential mapping inconsistencies and arrived at a set of 204 bilaterian-specific genes. As a final validation step, we blasted at NCBI (non-redundant GenBank version from May 24, 2017) all human or *D. melanogaster* orthologs, which are present in the 204 bilaterian-specific orthogroups, against non-bilaterian metazoans (Metazoa excluding Bilateria and Mesozoa). A reciprocal best hit analysis of the blast results indicated that 47 genes, corresponding to 47 orthogroups, might contain orthologs in non-bilaterian species although our orthology prediction pipeline did not detect them. As substantial work is required to confirm or reject these potentially false-positive orthogroups, we removed them from the list and arrived at a final number of 157 orthogroups. These 157 orthogroups represent a minimal set of high-confidence bilaterian-specific orthogroups which is free of most errors present in other orthology databases.

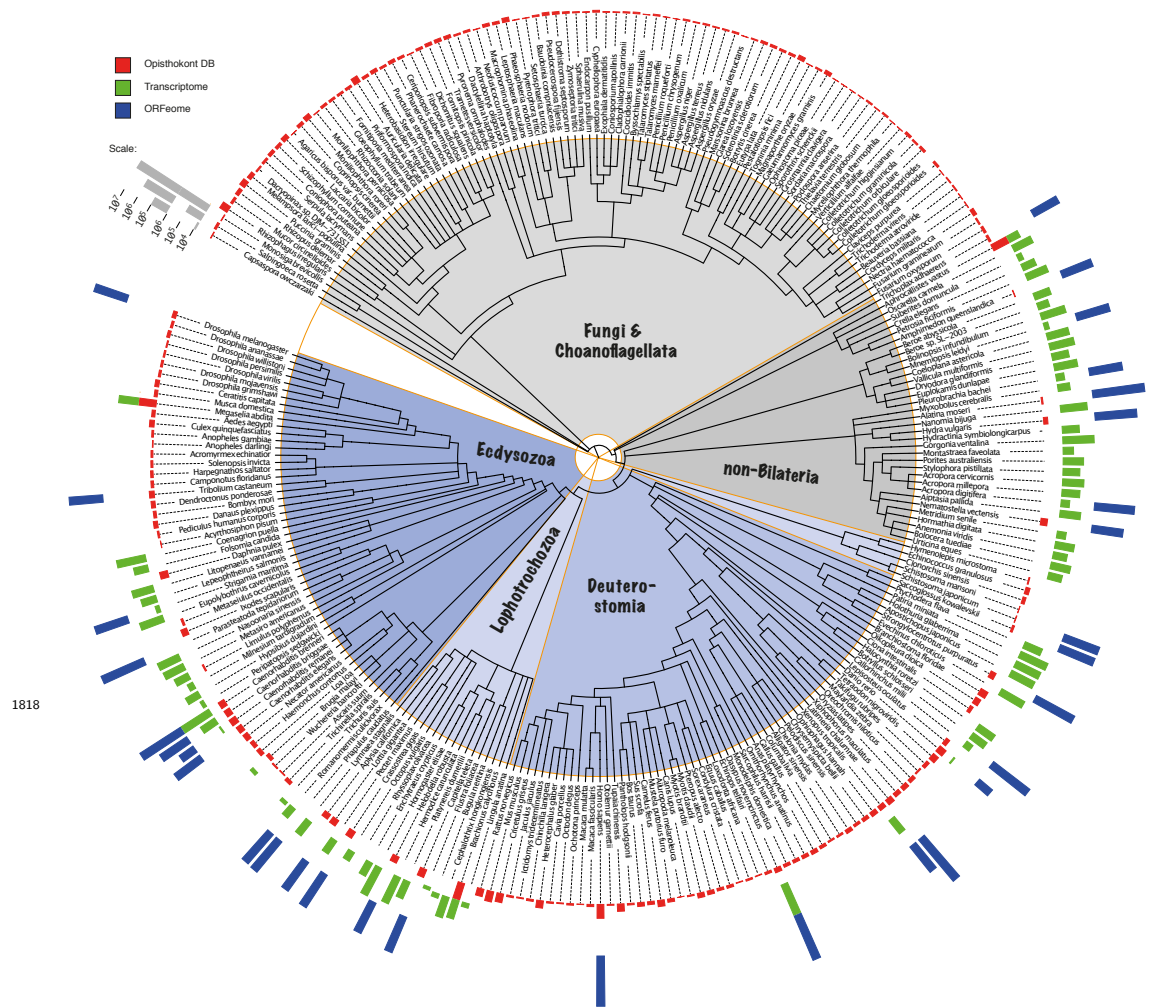


Figure 1-Figure supplement 1. Phylogenetic distribution of the BigWenDB. The amount of sequence data populating the BigWenDB is shown together with its phylogenetic distribution. The coloured bars at the perimeter (red, green, blue) document the contribution of three different sequence sources to the dataset (bar height proportional to the number of sequences, see ruler at top left): (1) Sequences from 204 opisthokonts (animals, choanoflagellates, and fungi) with >8,000 entries in the NCBI database (downloaded on May 25, 2015; coloured in red). (2) Sequences derived from the transcriptomes of 64 species under-represented at NCBI (e. g. non-bilaterian animals, lophotrochozoans, and representatives of additional phyla; green). (3) ORFs derived from the genome sequences of 25 representative metazoans (blue), including eight non-bilaterian species. In total, 124,031,501 sequences from 273 species cover the eukaryotic tree of life in the most comprehensive way so far (see text for details). Phylogenetic relationships after NCBI taxonomy.

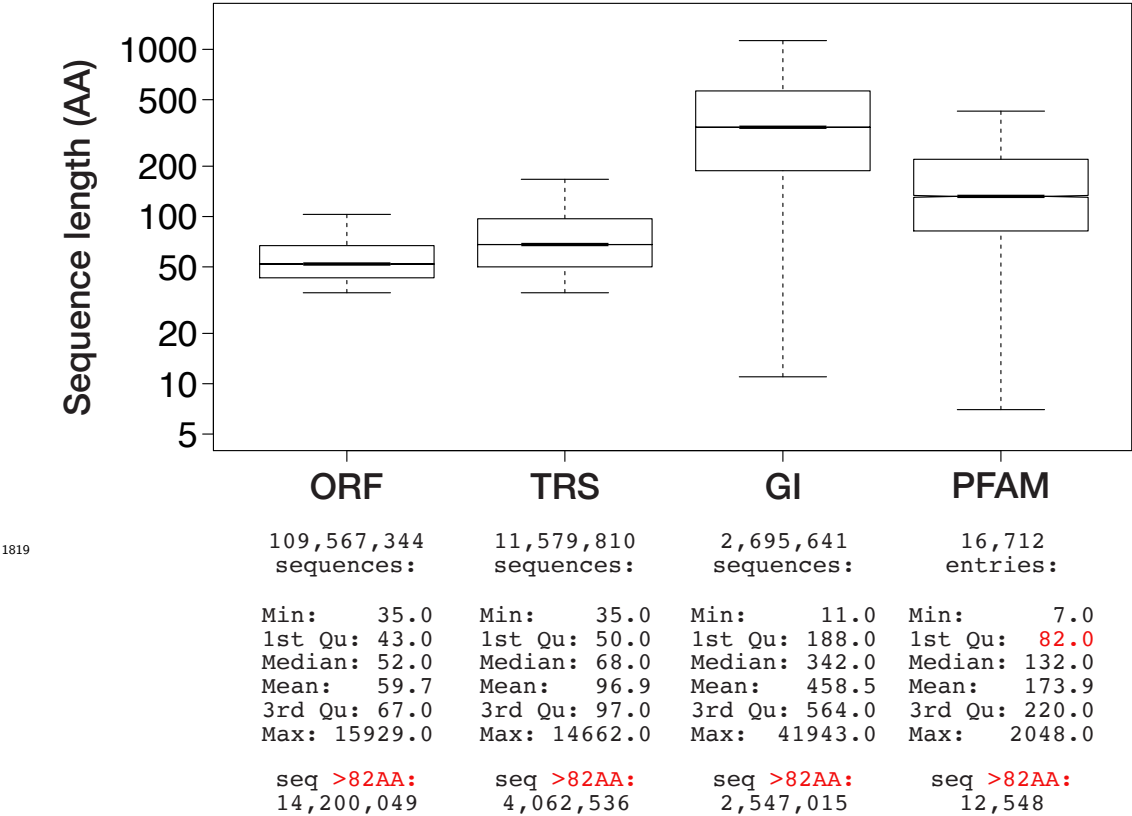


Figure 1-Figure supplement 2. Size distribution of three sequence data types present in the BigWenDB. Boxplots show the size distribution of genomic ORFs (ORF), transcriptomic ORFs (TRS), and NCBI sequences (GI) in comparison to the average size of protein domains collected in the PFAM database V31.0 (March 2017; 16,712 entries). Data points outside 1.5x the interquartile range are omitted for clarity. y-axis is in logarithmic scale. Corresponding sequence number and summary statistics are shown below each boxplot. The lower border (1st quartile) of the PFAM box is marked in red, together with the number of sequences per data type that surpass this size threshold.

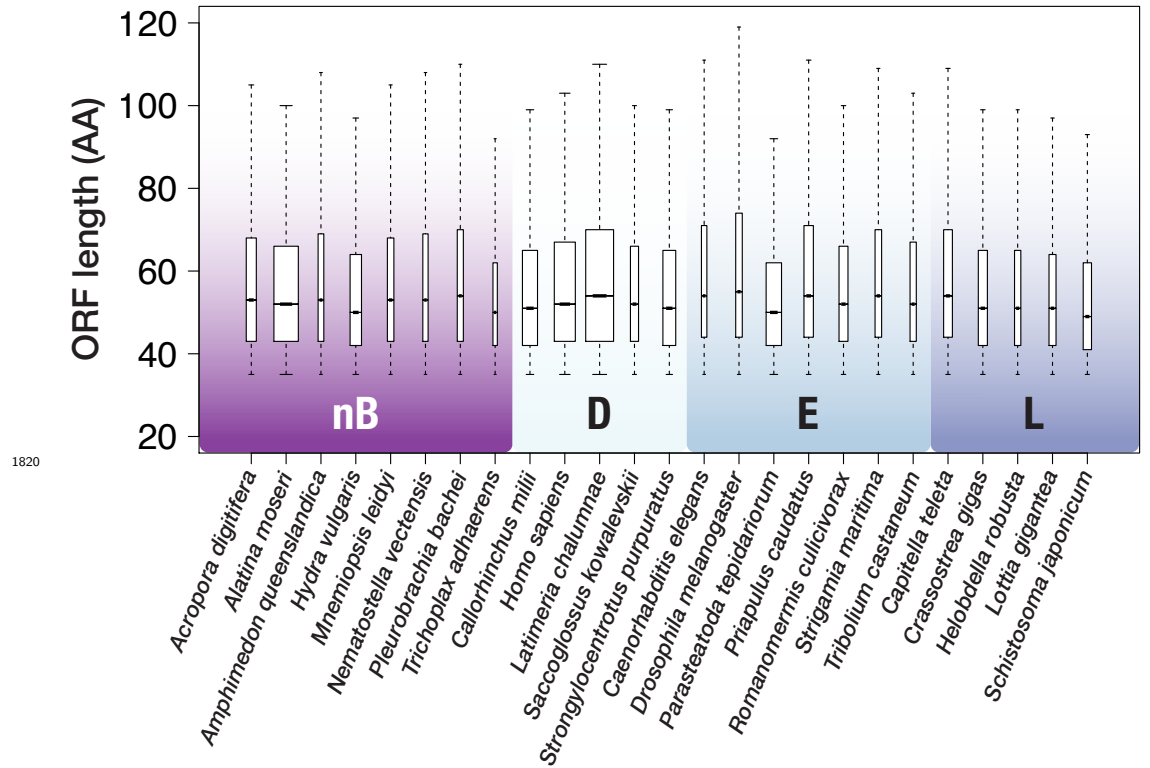


Figure 1-Figure supplement 3. ORF size distribution for 25 species with genomic data. Outliers (above whiskers) are omitted for clarity. Whiskers extend to 1.5x the interquartile range (default in R). Box width is proportional to the square root of the sequence number. nB = non-bilaterian Metazoa; D = Deuterostomia; E = Ecdysozoa; L = Lophotrochozoa.

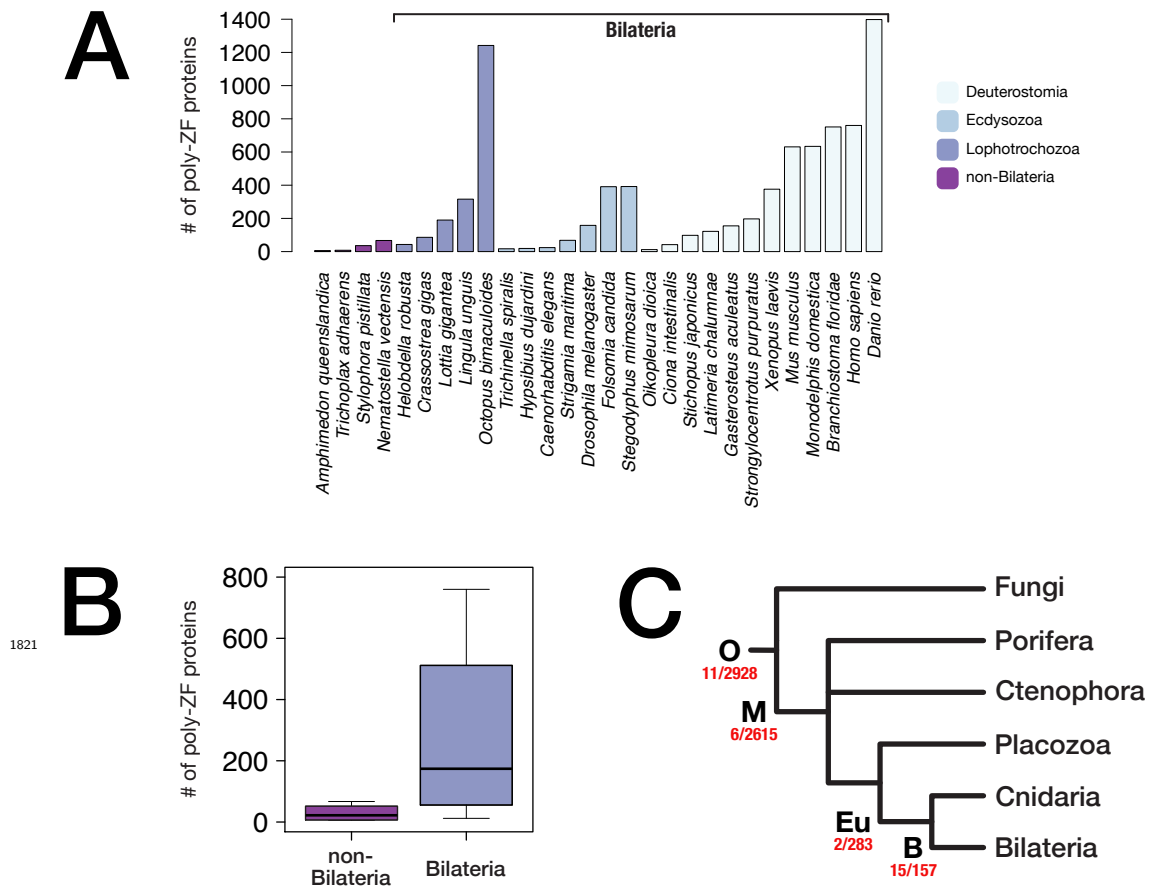


Figure 2—Figure supplement 1. Metazoan poly-zinc finger transcription factor repertoire and evolution. **A:** Reference proteomes of 28 representative metazoans (downloaded from uniprot.org) were scanned for the presence of C_2H_2 zinc finger proteins. For each species, the number of proteins with ≥ 6 domains is plotted. **B:** Boxplot representation of the number of poly-ZF proteins per genome in non-bilaterian Metazoa (4 species) vs. Bilateria (24 species) using scanning results of panel A. **C:** Evolutionary origin of poly-ZF proteins. On the basis of our orthology database (BigWenDB), we inferred lineage-specific orthogroups for four lineages, opisthokonts (O), metazoans (M), eumetazoans (Eu), and bilaterians (B), and analysed those orthogroups for the presence of C_2H_2 poly-ZF proteins with ≥ 6 domains. The number of such proteins vs. the total number of lineage-specific orthogroups is displayed in red under each node. «O» indicates origin in the ancestor of opisthokonts or earlier.

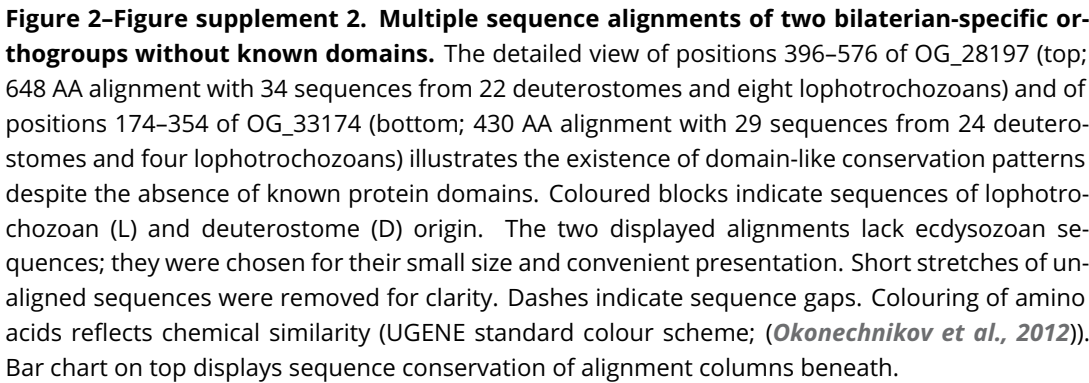
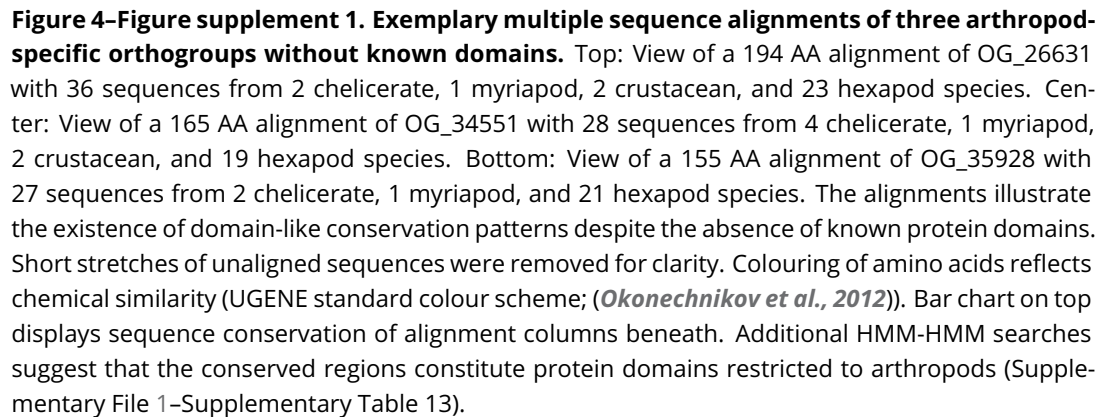




Figure 2–Figure supplement 4. Multiple sequence alignment of OG_8220, another bilaterian-specific orthogroup without known domains. View of a 234 AA alignment with 135 sequences from 22 deuterostomes, eight ecdysozoans, and nine lophotrochozoans, illustrating the existence of domain-like conservation patterns despite the absence of known protein domains. Short stretches of unaligned sequences were removed for clarity. Colouring of amino acids reflects chemical similarity (UGENE standard colour scheme; (*Okonechnikov et al., 2012*)). Bar chart on top displays sequence conservation of alignment columns beneath. Sequences are ordered according to their origin.



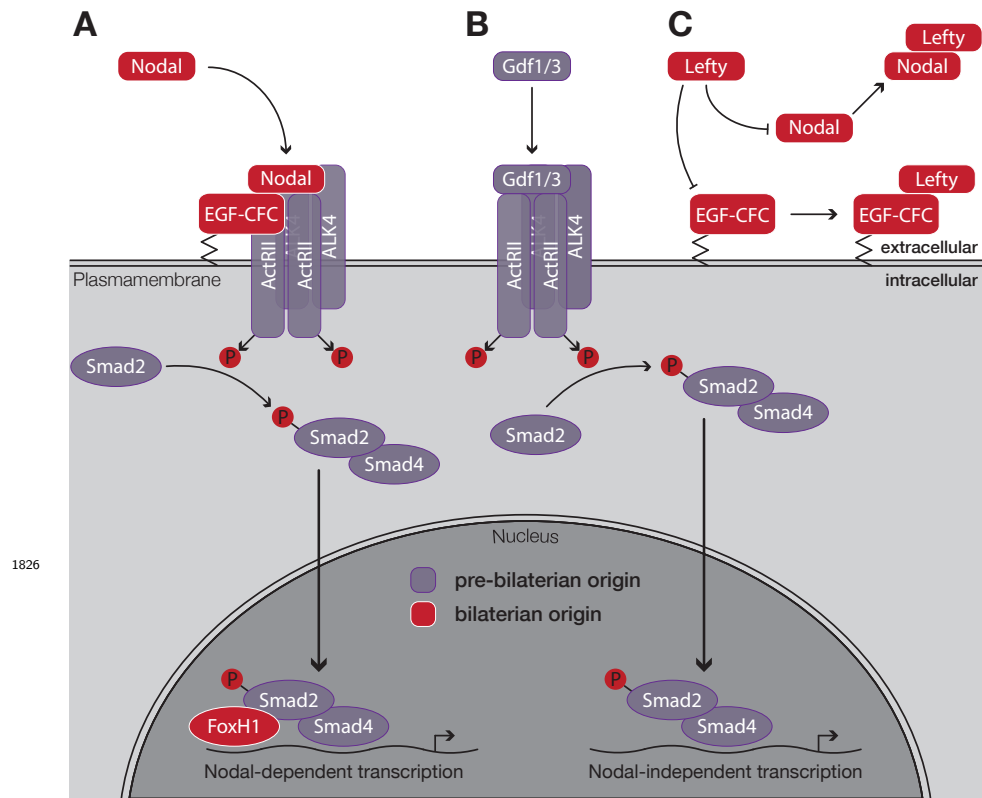


Figure 5—Figure supplement 1. Schematic outline of the Nodal signalling pathway in vertebrates. **A:** Nodal binds to its cell surface receptor in the presence of the co-receptor EGF-CFC, activating the resulting complex. After phosphorylation, the Smad2/Smad4 complex translocates to the nucleus. Upon binding of the transcription factor FoxH1, transcription of Nodal target genes is initiated. **B:** Nodal-independent transcription via the same pathway does not require the co-receptor EGF-CFC or the transcription factor FoxH1. **C:** Lefty antagonizes Nodal function by blocking either its co-receptor, EGF-CFC, or by directly binding to Nodal. Factors that evolved in the ancestor of bilaterians are displayed in red, all other factors evolved in the ancestor of eumetazoans or earlier. Figure modified after *Shen (2007)*.



Figure 5-Figure supplement 2. Bilateralian-specific distribution of the Nodal pathway components Nodal and Lefty. Maximum likelihood phylogeny of selected bilaterian Lefty and Nodal proteins. The corresponding multiple sequence alignment consists of 24 sequences with 446 columns and 29.01 % gaps and undetermined characters. The sequences correspond to OG_11821 (Lefty) and OG_12210 (Nodal) of the original clustering plus several additional candidate sequences from public repositories (red dots). Blue dots highlight whether a sequence is derived from transcriptomic (light blue) or genomic ORF data (dark blue). All other sequences can be accessed at NCBI with the gene identifiers given as branch labels. Blue triangles identify previously described Lefty and Nodal reference sequences. Bootstrap values below 50 % are removed for clarity. There are three Nodal-related genes in teleosts, cyclops, squint, and southpaw, as a result of lineage-specific duplications (*Fan and Dougan, 2007*).

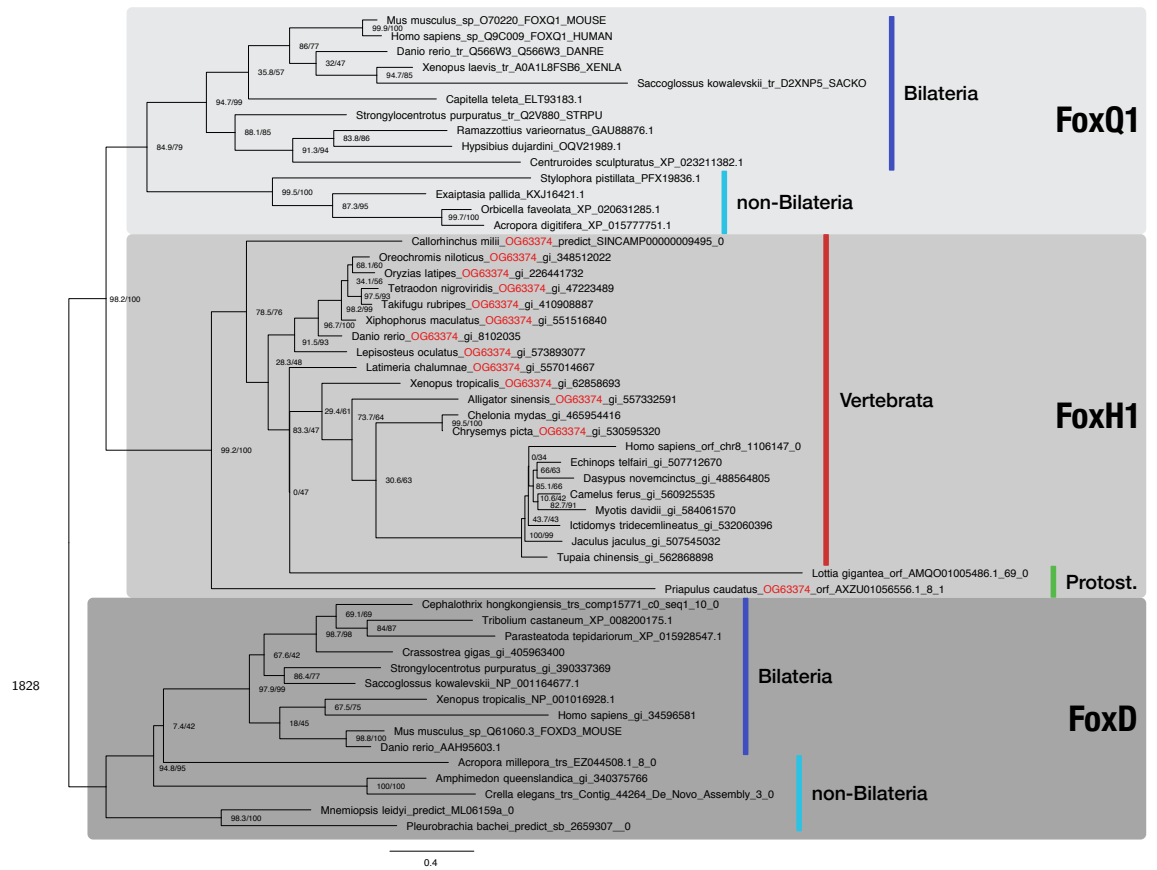


Figure 5–Figure supplement 3. Bilaterian-specific distribution of the Nodal pathway component FoxH1. Maximum likelihood phylogeny of selected metazoan Fox genes. The multiple sequence alignment consists of 52 sequences aligned over 315 positions (proportion of gaps and undetermined characters: 25.07 %). It is generated from OG_36001 (FoxH1), OG_63374 (RBH with OG_36001; orthogroup ID labeled in red), and representative sequences of OG_3972 (FoxD4 as outgroup; third-best hit of OG_36001 in HMM-HMM searches, see Supplementary File 1–Supplementary Table 14) of the original clustering. Selected FoxQ1 proteins were used as outgroup as FoxQ1 resembled the closest relative of FoxH1 proteins in other studies (Yu et al., 2008; Fritzenwanker et al., 2014). Vertebrate and protostomian FoxH1 sequences are decorated with a red and green bar, respectively. Sequences derived from genomic and transcriptomic ORFs are labelled with «|orf_», «|trs_», or «|predict_». All other sequences can be accessed at NCBI with the given identifiers. Branch labels correspond to the results of SH-aLRT (Shimodaira–Hasegawa-like approximate likelihood ratio test, left) and UFboot (ultrafast bootstrap approximation, right) as implemented in IQ-TREE (Nguyen et al., 2015).

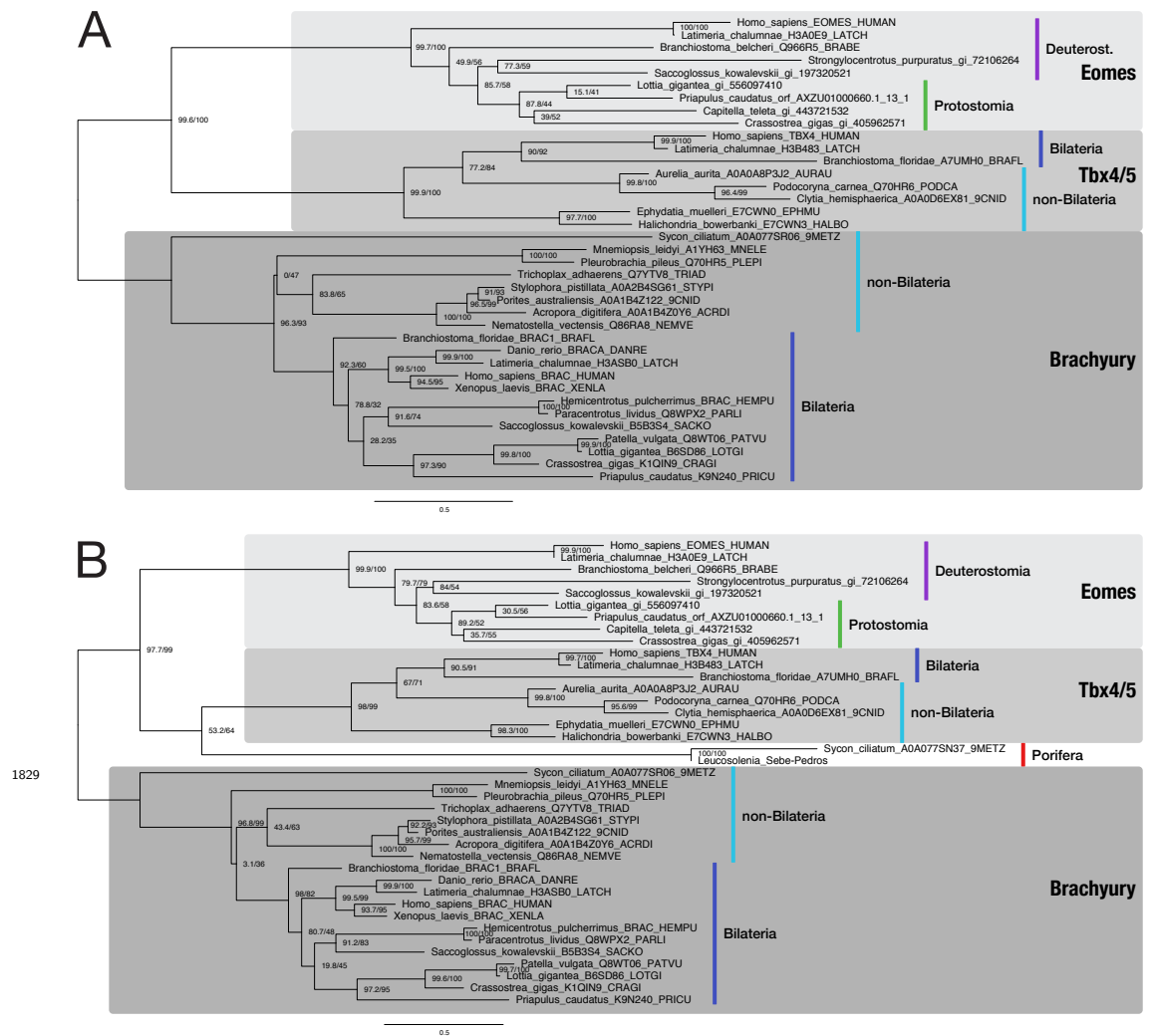


Figure 5—Figure supplement 4. Bilaterian-specific distribution of the Nodal pathway component Eomesodermin. **A:** Maximum likelihood phylogeny of selected poriferan and bilaterian Eomesodermin sequences. The multiple sequence alignment consists of 37 sequences aligned over 434 positions (proportion of gaps and undetermined characters: 22.80 %). Sequences were downloaded from uniprot.org or taken from NCBI ($|g_*$). TBX4 and Brachyury sequences serve as outgroups because they are most closely related to the Eomes family according to *Sebé-Pedrós et al. (2013)* and HMM-HMM searches (Supplementary File 1–Supplementary Table 14). A phylogenetic analysis with an identical dataset, including the two poriferan Eomes candidates (highlighted in red; from *Sebé-Pedrós et al. (2013)*), is presented in panel **B** (39 sequences aligned over 435 positions; proportion of gaps and undetermined characters: 23.64 %). Branch labels correspond to the results of SH-aLRT (Shimodaira–Hasegawa-like approximate likelihood ratio test, left) and UFBoot (ultrafast bootstrap approximation, right) as implemented in IQ-TREE (*Nguyen et al., 2015*). Tree topology and corresponding bootstrap values do not clearly assign the poriferan sequences to the Eomes family of T box proteins.

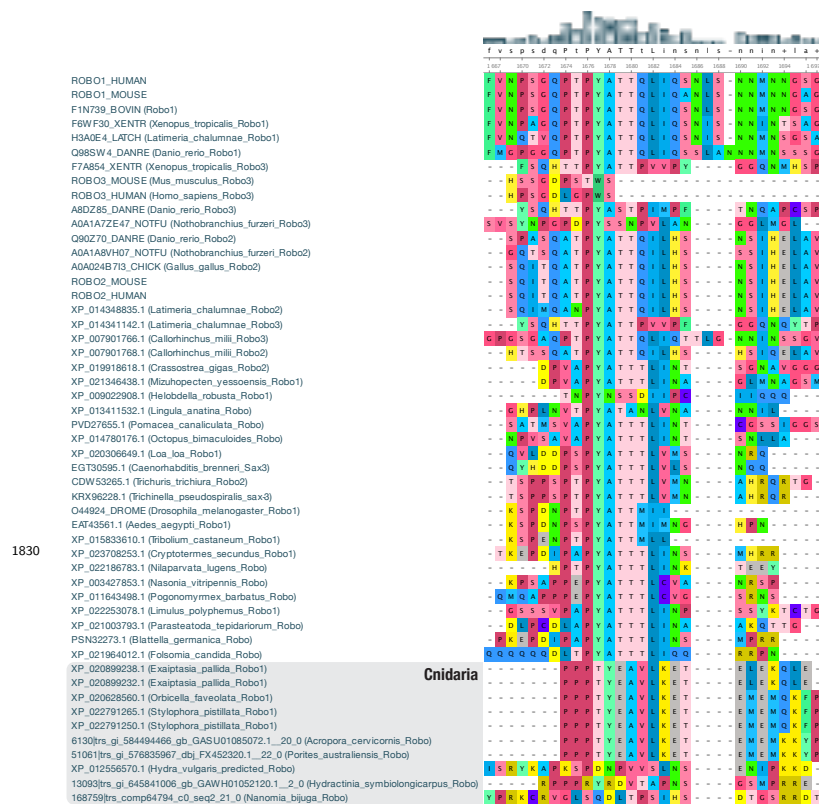


Figure 6-Figure supplement 1. Change of the conserved cytoplasmic motif CC1 in cnidarian Robo-like proteins. Multiple sequence alignment of 41 bilaterian and ten cnidarian (bottom) Robo proteins. A fragment of the full alignment is shown (AA 1667–1697), centering on the conserved cytoplasmic motif CC1 (corresponding to sequence «TPYATTQLI» of human Robo1). Colouring of amino acids reflects chemical similarity (UGENE standard colour scheme; (*Okonechnikov et al., 2012*)). Bar chart on top displays sequence conservation of alignment columns beneath. Despite the presence of a potential tyrosine phosphorylation site (Y), the CC1 motif is not conserved in cnidarian Robo-like proteins.

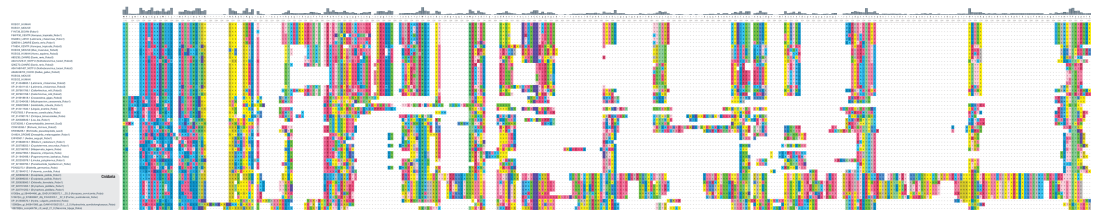


Figure 6-Figure supplement 2. Cnidarian Robo-like proteins display structural alterations. Multiple sequence alignment of 41 bilaterian and ten cnidarian (bottom) Robo proteins. A fragment of the full alignment is shown (AA 1271–1617), starting with the transmembrane region (blue part on the left). Colouring of amino acids reflects chemical similarity (UGENE standard colour scheme; (*Okonechnikov et al., 2012*)). Bar chart on top displays sequence conservation of alignment columns beneath. Cnidarian Robo-like proteins possess insertions and deletions relative to bilaterian Robos, especially at the beginning of the cytoplasmic part.

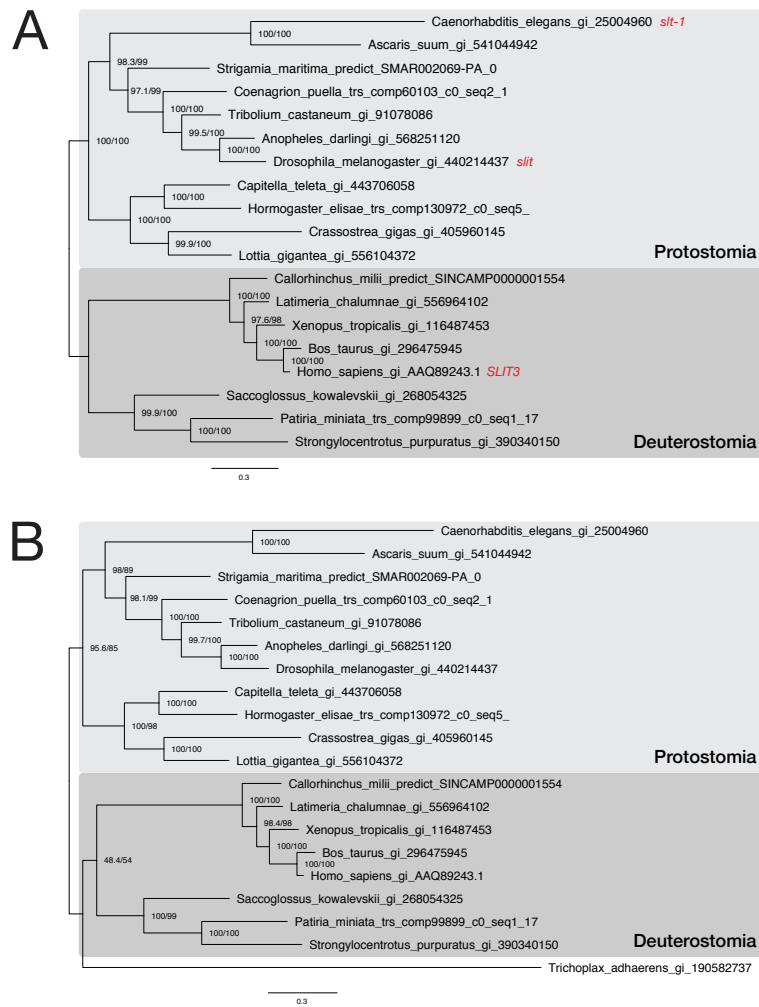


Figure 6–Figure supplement 3. Phylogenetic analysis of a putative *Trichoplax adhaerens* Slit protein. Two maximum likelihood phylogenies of representative bilaterian Slit sequences. Sequences were downloaded from NCBI or extracted from the corresponding Slit orthogroup OG_5717. In subfigure A, the multiple sequence alignment consists of 19 sequences aligned over 1,800 positions (proportion of gaps and undetermined characters: 24.15 %). In B, a single protein from the placozoan *Trichoplax adhaerens* was added to the dataset, generating an alignment of 20 sequences over 1,865 positions (proportion of gaps and undetermined characters: 26.73 %). Branch labels correspond to the results of SH-aLRT (Shimodaira–Hasegawa-like approximate likelihood ratio test, left) and UFboot (ultrafast bootstrap approximation, right) as implemented in IQ-TREE (Nguyen *et al.*, 2015). Tree topology and corresponding bootstrap values are compatible with assigning the placozoan sequence to the Slit protein family.

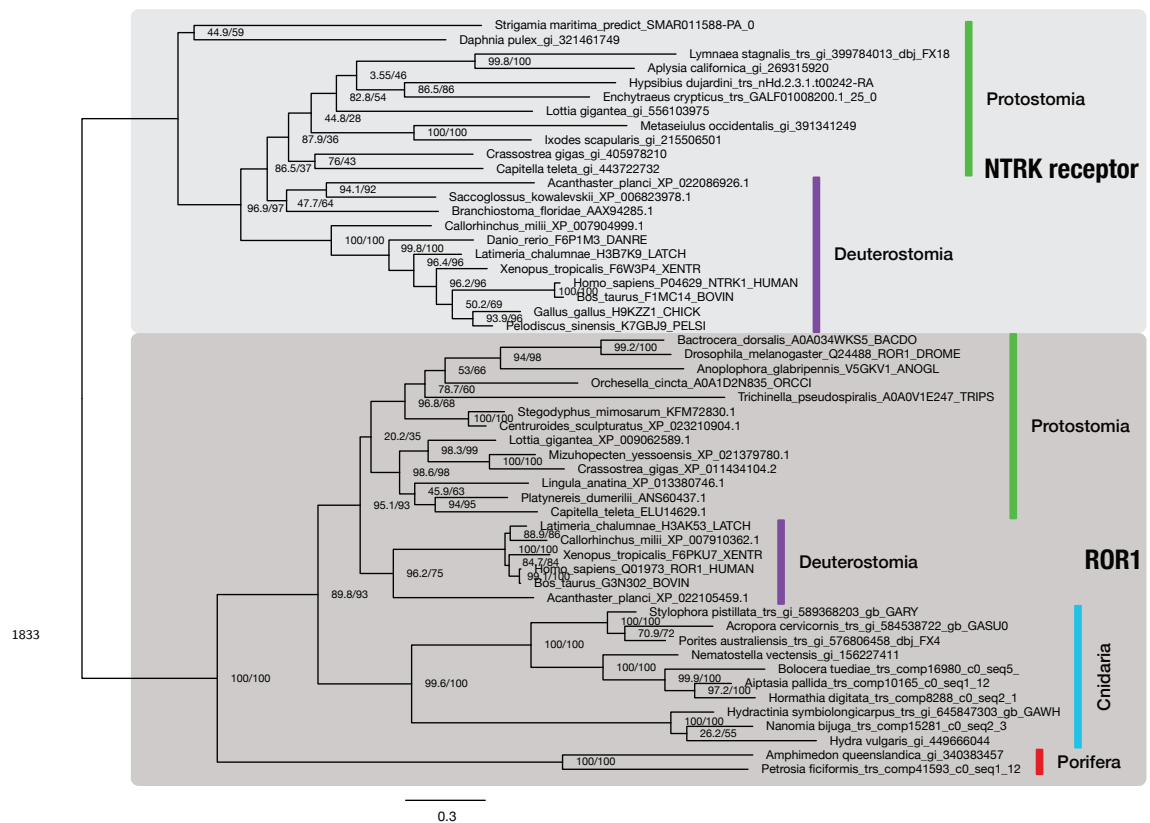


Figure 7—Figure supplement 1. The NTRK neurotrophin receptor is restricted to bilaterians. Maximum likelihood phylogenetic analysis of 53 metazoan NTRK and ROR1 sequences (out-group), aligned over 602 AA. Proportion of gaps and completely undetermined characters in the corresponding alignment: 16.84 %. Sequences were collected from different sources: NTRK receptor sequences from protostomes are derived from OG_8965-1.4 of the 1.4 clustering, an orthogroup containing RTKs only (Supplementary File 1—Supplementary Table 24). Deuterostomian NTRK sequences were collected at www.uniprot.org. Non-bilaterian ROR1 sequences were obtained from OG_6493-1.4, the ROR1-specific orthogroup of the 1.4 clustering (Supplementary File 1—Supplementary Table 24), while most bilaterian ROR1 sequences were downloaded from www.uniprot.org. Branch labels correspond to the results of SH-aLRT (Shimodaira–Hasegawa-like approximate likelihood ratio test, left) and UFBoot (ultrafast bootstrap approximation, right) as implemented in IQ-TREE (Nguyen *et al.*, 2015).